O'REILLY®

Compliments of TalkingData

Implementing a Smart Data Platform

How Enterprises Survive in the Era of Smart Data

Yifei Lin & Wenfeng Xiao


TalkingData
16.51
8.23
9.94
9.49
17.71
12.48
11.11
17.28
21.71
24.45
16.09
20.45
9.94
26.61
34.54
17.34
29.94
China's leading
SMART DATA PLATFORM
Go Beyond With The Heart And Mind Of Data

Implementing a Smart Data Platform

How Enterprises Survive in the Era of Smart Data

Yifei Lin and Wenfeng Xiao

Beijing • Boston • Farnham • Sebastopol • Tokyo O'REILLY®

Implementing a Smart Data Platform
by Yifei Lin and Wenfeng Xiao

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

Editor: Nicole Tache
Production Editor: Melanie Yarbrough
Copyeditor: Jasmine Kwityn
Proofreader: Charles Roumeliotis

Interior Designer: David Futato
Cover Designer: Karen Montgomery
Illustrator: Rebecca Demarest

May 2017: First Edition

Revision History for the First Edition
2017-05-10: First Release

978-1-491-98346-1

[LSI]

Table of Contents

CHAPTER 1

The Advent of the Smart Data Era

The data we collect has experienced exponential growth, whether we get it through our PCs, mobile devices, or the IoT, or from tools for ecommerce or social networking. According to the IDC Report, global data volume reached 8 ZB (or 8 billion TB) in 2015 and is expected to reach 35 ZB in 2020, with an annual increase of nearly 40%. And according to TalkingData, in 2016 China was home to 1.3 billion smartphone users, accounting for tens of millions of wearable devices such as smart watches and over 8 billion sensors of different kinds. Smart devices can be seen nearly everywhere and generate data of various dimensions—anytime, anywhere.

Data accumulation has created favorable conditions for the development of artificial intelligence (AI). The training of machines with a huge amount of data may generate more powerful AI. For example, the game of Go (or "Weiqi" in Chinese) has been traditionally viewed as one of the most challenging games due to its complicated tactics. In 2016, Google's program AlphaGo (with access to 30 million distributed data points and improved algorithms, accumulated by users after they played Go hundreds of thousands of times) defeated world Go champion Li Shishi, proving its No.1 Go-playing ability. In the previous two years, AI also witnessed explosive growth and application in the fields of finance, transport, medicine, education, industry, and more. It's clear that the data accumulated by mankind has been used to produce new intelligence, which could aid our work, reduce costs, and improve efficiency. According to a CB Insights report, investment funds of global AI startups also had exponential growth during 2010 to 2015.

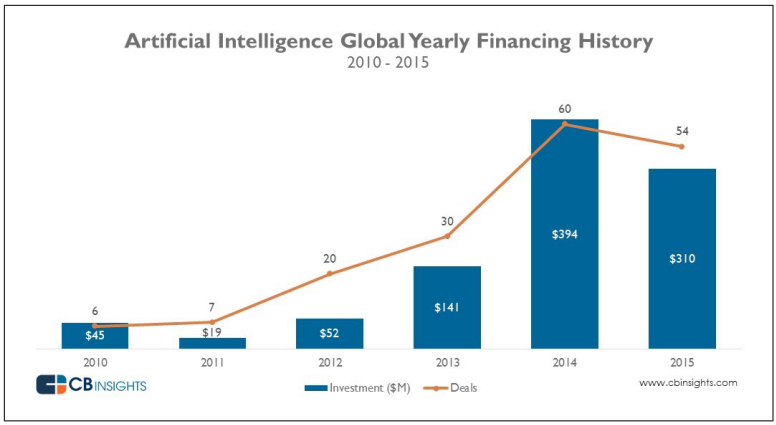

Figure 1-1. Artificial intelligence global yearly financing history, 2010–2015, in millions of dollars (source: CB Insights)

Data accumulation and the development of AI promote and complement each other. Andrew NG, AI expert and VP & Chief Scientist of Baidu, said in a *Wired* article, "To draw an analogy, data is like the fuel for a rocket. We need both a big engine (algorithm) and plenty of fuel (data) in order to enable the rocket (AI) to be launched." Also, AI has brought us more application contexts such as chatting robots and autonomous vehicles, which are generating new data.

And now data is becoming not only bigger but also smarter and more useful. We have entered the smart data era.

Three Elements of the Smart Data Era: Data, AI, and Human Wisdom

Data accumulation can enable deeper insights and help us to gain more experience and wisdom. For example, through further analysis on mobile phone users' behaviors, enterprises can gain more understanding of their clients, including their preferences and consuming habits, so as to gain more marketing opportunities. Additionally, AI in itself requires the involvement of human wisdom so as to guide the orientation of AI and increase its efficiency. For example, AlphaGo needs to fight against professionals in the game of Go so as to continuously enhance its Go-playing ability with the aid of human wisdom.

Without the continuous intervention of human wisdom, the addition of AI to data will lose some of its value and even become ineffective. Conversely, without AI, it is a challenge for humans alone to deal with such complicated and rapidly changed data. Also, without data, it would be impossible for AI to exist and the accumulation of human wisdom would also slow down. Data, AI, and human wisdom facilitate each other and form a forward loop.

For example, in the field of context awareness, the movements and gestures of mobile phone users (including walking, riding, driving, etc.) may be judged by using AI algorithms with the phones' sensor data. If any judgment is not accurate enough, data should be sorted and enhanced by human intervention and algorithms should be optimized until the result is acceptable. Also, mobile phones capable of context awareness may provide application developers more contexts and experience, such as body-building (i.e., gestures need to be captured and the frequency/number of steps or even the place needs to be judged in order to obtain more accurate data of users' status), financial risk control, logistics management, and entertainment. Accordingly, more data would be generated. This new data may allow human wisdom to grow quickly and AI to become more powerful. For example, it is discovered through context-awareness data that most users keep their mobile phones in their hands when they are using apps. Thus, does a non-handheld application context—such as fraudulent app rating, done on non-handheld mobile phones—mean even greater financial risk?

The three elements of the smart data era have generated incredible value in their combined and independent actions. Enterprises that adapt to the new era would be able to restructure their infrastructure using data, AI, and human wisdom and accelerate the process of exploring and realizing commercial value so as to stand out in fierce competition. Those enterprises with slow actions would be at a loss when they are faced with scattered and complicated data and gradually lose their competitiveness. There is no way for them to share the greatest benefit (i.e., value). Nevertheless, the shock of a new era is independent of enterprise scale or industry.

In this report, we are going to list the challenges for enterprises during the smart data age and analyze their causes. With over five years of industrial service experience, TalkingData has helped enterprises find solutions to cope with the challenges of data, and to efficiently explore the business value of data. We introduce the concept of

SmartDP along with the three basic capabilities that SmartDP should possess: data management, data science, and data engineering. Meanwhile, we also introduce the SmartDP referential framework, and detail the functions of each layer. Finally, we will take a look at how SmartDP is adopted in real scenarios to enhance our understanding of smart data.

CHAPTER 2

Challenges of the Smart Data Era for Enterprises

In the smart data era, enterprises should transform themselves from traditional product- and technology-driven enterprises into data-driven ones. Different from traditional enterprises, data-driven enterprises are characterized by the following aspects:

- Data is regarded as an important asset for management.
- Specific data applications are used to solve business problems. (These applications are linked to the current data systems of enterprises. Meanwhile, enterprise data—both self-owned and other business-related data—are called).
- Specialized and structured data teams are set up inside the enterprises (problems are not solved by outsourcing).
- A data-driven culture is built.

During the transition to becoming a data-driven entity, traditional enterprises are severely challenged by business digitalization and data capitalization. A huge amount of data is not acquired in an effective manner due to the lack of business digitalization. For example, users' click event data on websites, the interaction data of app users, user subscription and browsing data on WeChat public platforms, customer visit data of offline stores, and other business-related data may not be acquired or used. Nowadays, the prevailing mobile phones (e.g., iPhone, Samsung Galaxy, etc.) are generally equipped with 15 or more sensors, including ambient light condi-

tion perception, acceleration, terrestrial magnetism, gyroscope, distance, pressure, RGB light, temperature, humidity, Hall coefficient, heartbeat and fingerprint, and more. If all sensors are activated, each mobile phone could acquire up to 1GB of data per day. Although this data can truly present the contexts of mobile users, most is abandoned.

With both the scale and dimensions of data rapidly increasing, enterprises are unable to effectively prepare and gain insight from data, making it hard for them to support business policymaking. According to a report of BCG (Boston Consulting Group) in 2015, only 34% of the data generated by financial institutions (with a relatively higher degree of IT support) was actually used. And according to a survey report of Experian Data Quality, in 2016 nearly 60% of American enterprises could not actively sense or deal with the issue of data quality and did not have fixed departments or roles responsible for managing data quality. There is clearly still a long way to go in terms of managing complicated data. If not effectively utilized, a large amount of data would not be asset-oriented and thus would not produce any value, which means huge costs for enterprises in turn.

Enterprises struggle with these challenges for a variety of reasons: some have no advanced technical platform, some are deficient in data management, some have not built standard data engineering systems, and some others simply lag behind in terms of their understanding of the value of data science. All these have hampered the transformation of traditional enterprises toward intelligent, data-driven ones. Let's look at each of these challenges more closely.

Challenges in Data Management

First, enterprises are faced with a series of challenges that need be solved by proper data management. These challenges include:

- Numerous internal systems and inconsistent data might cause confusion. Take gender, for example. It may differ in a CRM system (actual gender in the fundamental demographics), a marketing system (e.g., a husband may sometimes purchase female-oriented goods in order to send a gift to his wife), and a social networking system (e.g., unique sexual orientation). If

gender is purely regarded as a consistent attribute across systems, errors may occur.

- The descriptive information of data (metadata) is controlled by different people in different departments of an enterprise, and fails to be shared across channels. Even for the same data, the understanding how it may be different due to the possible existence of varying standards. For example, the HR Department of an enterprise would maintain a list of employees and their addresses (home addresses) but the Administration Department may update an address to send employee benefits for the holidays so that such benefits can be properly delivered. In such cases, "home addresses" are changed to "mailing addresses." However, both parties believe that the correct addresses have been given. Another example is ecommerce. For the number of ecommerce apps activated, the Marketing Department may believe that apps are activated after they are started for the first time but the Product Department may think that apps are activated once they are used to make a purchase for the first time.
- It is difficult to effectively integrate the data that is distributed on the enterprise's external platforms. For example, the data acquired by a WiFi probe installed in the store of an enterprise and the data accumulated on each third-party media platform (such as the WeChat public platform) may possibly supplement client data dimensions. However, the IDs used for client follow-up fail to be connected. As a result, the data of all platforms is unable to sync, thus greatly reducing the value of data.

Challenges in Data Engineering

Second, enterprises encounter challenges when data and the current business flow don't form a complete value chain. In such a case, data engineering is required to solve the issue. These challenges include:

- Lack of explicit data standards and specifications. Each department or system gives different definitions or descriptions of the same data and acquires data of varying quality, or even misses some data in acquisition, which burdens the data processing later.
- Lack of explicit definitions about job functions and engineering of data. Data management work is assigned to people at ran-

dom, typically IT personnel, data architects, data analysts, or data scientists. Also, there are instances when no specific rights and responsibilities are designated to those working with data. As a result, it becomes difficult to conduct continuous data management operation and form a closed loop.

- Increasing data application contexts and the data processed by various data applications leads to redundant and ineffective data preparation and analysis, thus impacting the efficiency of delivering the data applications.

Challenges in Data Science

Third, shifting practical issues to automatic decisions that can be supported by data also introduces challenges, which need to be solved by data science. These challenges include:

- Shortage in data science professionals. It seems quite difficult to apply the most cutting-edge technologies of data science as there are not many talents in the field of data science. McKinsey estimated that 190,000 additional data scientists are needed in the United States by 2018, and that figure would be even bigger in China.
- If the quality of data is unstable, it is difficult to see its value, even if the algorithms used on that data are in working order. According to an EDQ report, the biggest factors that affect data quality include incomplete or lost data, obsolete information, repeated data, inconsistent data, and flawed data (e.g., containing spelling errors). In order to solve these problems, systematic considerations should be made. Thus, it would be difficult for these problems to be solved only by stopgap measures.
- Enterprises are too eager for quick success and instant benefits to make long-term investments in the data field. Data science is never a cure-all and it is difficult for it to solve all problems in one stroke. In most cases, continuous investment is required. Gradual improvements should be made with algorithm optimization and iterative models that cover each link of data engineering, including data acquisition, organization, analysis, and action. Take the marketing and launching of applications for example. The audience for one round of the launch should be adjusted according to the results of the previous round. The

launch process can be improved only after several rounds of iteration.

Challenges in Technical Platform

Finally, the data management, data engineering, and data science teams also present a challenge to the technical platform. The challenges to the platform include:

- Increasing scale and dimensions of data. In the past, the data acquired by enterprises was mainly derived from emails, web pages, call centers, and so on. Currently, data sources also include mobile phone applications, sensors (such as iBeacon), social media, VR/AR devices, automobiles, and smart home appliances. The data being obtained by enterprises is becoming more and more varied, and helps these organizations capture a huge amount of data of various dimensions.
- Increasing data sources and types. In addition to traditional structured data, semi-structured data (such as JSON), non-structured data (such as videos, images, and texts) and flow type data (such as click blogs on websites) should also be processed. In addition to the enterprise's own data stored in internal CRM systems and public platforms such as WeChat, third-party data purchased by enterprises from the data trading market may also need to be processed.
- Continuously changing data formats. This is the most common challenge in the current data ecology. For example, an upstream data provider may fail to notify all downstream data providers when it adjusts a data format. Additionally, a change in data dimensions upon acquisition may often cause challenges. For instance, a particular sensor might be added to a newly released smart mobile phone, which may require the addition of new fields in the data format collected.
- As enterprises gradually shift their demands for data analytics from simple presentation to backend business support, there is an increasingly higher demand for real-time performance of the data platform. For example, many results of real-time data statistics now show changes in the real-time customer flow of apps or offline stores and tell us when there are the most visits or which public platform or store is the most active. Also, such

results can be used to analyze the flow or number of clients at individual hours of a day. This is of great significance for the time management and resource allocation of websites.

CHAPTER 3

The Advent of Smart Enterprises and SmartDP

Despite the difficulty of transition in the smart data era, many emerging enterprises rose above others and enhanced their competitiveness with data, which shocked traditional enterprises in all fields. According to the Mobile Internet Report 2016 issued by A16Z, the data giants represented by GAFA (Google, Amazon, Facebook, and Apple) have accumulated competitive advantages in the fields of data and technology and earned more than three times the revenue of Wintel (Microsoft and Intel) on an annual basis. In turn, they are changing the forms and modes of traditional industries through data and technology, including retail, media distribution, automotive, and so on.

These new pioneers share something in common: they have implemented a data-driven business model and a sophisticated data asset management system. Furthermore, they are able to drive contextual applications by using data, as well as explore and convert commercial value in an efficient manner. Such enterprises that have built a data-driven culture are called smart enterprises. Characteristics of smart enterprises include the following:

- Their flexible technical platforms and data science capacity can sufficiently support huge data scale, large data dimensions, complicated data types, and flexible data formats. These platforms also enable quick insights from data, which increases the efficiency of various data application contexts.

- Their unified data management strategy can be used to manage data views that are consistent across enterprises, efficiently gather data (including self-owned and third-party data), and also efficiently output data and data services.
- Their end-to-end data engineering capacity can support data management for the business and help form a closed loop that continuously optimizes business operations.

Smart enterprises are the companies that are armed with these three capabilities.

In order to become data-driven, smart enterprises need a new platform to support them, a platform that promotes an environment that is focused on data. This platform is called SmartDP (smart data platform). SmartDP refers to a platform that explores the commercial value of data based on smart data applications, and enables proper data management, data engineering, and data science.

Comprised of a set of modern data solutions, SmartDP helps enterprises build an end-to-end closed data loop, from data acquisition to decision to action, in order to provide the capacity for flexible data insight and data value mining as well as flexible and scalable support for contextual data applications. As we'll see later in this report, adopting SmartDP can improve enterprises' data management, data engineering, and data science capabilities. We'll now review each of these aspects in general terms.

CHAPTER 4

Data Management, Data Engineering, and Data Science Overview

Data Management

Data management refers to the process by which data is effectively acquired, stored, processed, and applied, aiming to bring the role of data into full play. In terms of business, data management includes metadata management, data quality management, and data security management.

Metadata Management

Metadata can help us to find and use data, and it constitutes the basis of data management.

Normally, metadata is divided into the following three types:

- *Technical metadata* refers to a description of a dataset from a technical perspective, mainly form and structure, including data type (such as text, JSON, and Avro) and data structure (such as field and field type).
- *Operational metadata* refers to a description of a dataset from the operation perspective, mainly data lineage and data summaries, including data sources, number of data records, and statistical distribution of numerical values for each field.

- *Business metadata* refers to a description of a dataset from the business point of view, mainly the significance of a dataset for business users, including business names, business descriptions, business labels, data-masking strategies.

Metadata management, as a whole, refers to the generation, monitoring, enrichment, deletion, and query of metadata.

Data Quality Management

Data quality is a description of whether the dataset is good or bad. Generally, data quality should be assessed for the following characteristics:

- *Integrity* refers to the integrity of data or metadata, including whether any field or any field content is missing (e.g., the home address only contains the street name, or no area code is included in the landline number).
- *Timeliness* or "freshness" refers to whether data is delayed too long from its generation to its availability and whether updates are sufficiently frequent. For example, real-time high-density data updates are necessary for the status monitoring of servers to ensure an alarm can be sent and dealt with in a timely manner in case of any problem to avoid more serious problems. To track the number of new mobile app users, Daily Active Users (DAUs) should be updated once a day in general cases. However, the increase in the number of new users is rarely studied.
- *Accuracy* refers to whether data is erroneous or abnormal—for example, incorrect phone numbers, having the wrong number of digits in an ID number, and using the wrong email format.
- *Consistency* involves both format (e.g., whether the telephone number conforms to MSISDN specifications) and logic across datasets. Sometimes, it may be OK from the point of view of a single dataset. However, problems would occur if two datasets are interconnected. For example, inconsistent gender data may appear in the internal system of an enterprise. The data may show male in the CRM system but female in the marketing system. Data should be further understood so as to adjust data descriptions and ensure data visitors are not confused.

Data quality management involves not only index description and monitoring of data integrity, timeliness, accuracy, and consistency but also the improvement of data quality by means of data organization.

Sometimes the problems of data quality are not so conspicuous and there is no way to make judgments only by statistical figures—in these cases, domain knowledge is required. For example, when the Tencent data team performed a statistical analysis of SVIP QQ users, it was discovered that the age group at 40 years old was the largest of such users, far more than the ages of 39 and 41. It was thus guessed that the group had an increased opportunity for online communication with their children or more free time. However, this did not align with the domain knowledge, which was not convincing. Further analysis revealed there were an inordinate number of users with a birthdate of January 1, 1970—the default birthdate set by the system—and that this is what had accounted for the high number of 40-year-old users (the study was conducted in 2010). Therefore, data operators should have a deep understanding of data and obtain the domain knowledge that is not known by others.

Data Security Management

Data security mainly refers to the protection of data access, use, and release processes, which includes the following:

- *Data access control* refers to the control of data access authority so that data can be accessed by the personnel with proper authorization.
- *Data audit* refers to the recording of all data operations by log or report so as to be traceable if needed.
- *Data mask* refers to the deletion of some data according to preset rules (especially the parts concerning privacy, such as personally recognizable data, personal private data, and sensitive business data) so as to protect data.
- *Data tokenization* refers to the substitution of some data content according to preset rules (especially sensitive data content) so as to protect data.

Data security management, therefore, entails the addition, deletion, modification, and monitoring of data, which aims to enable users to

access data in a convenient and efficient manner while ensuring data security.

Data Engineering

Most traditional enterprises are challenged by poor implementation of data acquisition, organization, analytics, and action procedures when they transform themselves for the smart era. Thus, it is urgent that enterprises build end-to-end data engineering capacity throughout their data acquisition, organization, analytics, and action procedures, so as to ensure a data- and procedure-driven business structure, rational data, and a closed-loop approach, and realize the transformation from further insight into commercial value of data. The search engine is the simplest example. After a search engine makes a user's interactive behavior data-driven, it can optimize the presentation of the search result so as to improve the user's searching experience and attract more users to it. This optimization is done according to duration of the user's stay, number of clicks, and other conditions. Additionally, it can generate more data for optimization. This is a closed loop of data, which can bring about continuous business optimization.

In the smart data era, due to the complexity of data and data application contexts, data engineering needs to integrate both AI and human wisdom to maximize its effectiveness. For example, a search engine aims to solve the issue of information ingestion after the surge in the volume of information on the internet. As tens of millions of web pages cannot be dealt with using manual URL classified navigation, algorithms must be used to index information and sort search results according to users' characteristics. In order to adapt to the increasingly complex web environment, Google has been gradually improving its search ranking intelligence, from the earliest PageRank algorithm, to Hummingbird in 2013 and the addition of the machine learning algorithm RankBrain as the third-most important sorting signal in 2015. There are over 200 sorting signals for the Google search engine; and variant signals or subsignals may be in the tens of thousands and are continuously changing. Normally, new sorting signals need to be discovered, analyzed, and evaluated by humans in order to determine their effects on the sorting results. Thus, even if there are powerful algorithms and massive data, human wisdom is absolutely necessary and undertakes a key role in efficient data engineering.

Implementation Flow of Data Engineering

In terms of implementation, data engineering normally includes data acquisition, organization, analytics, and action, which form a closed loop of data (see Figure 4-1).

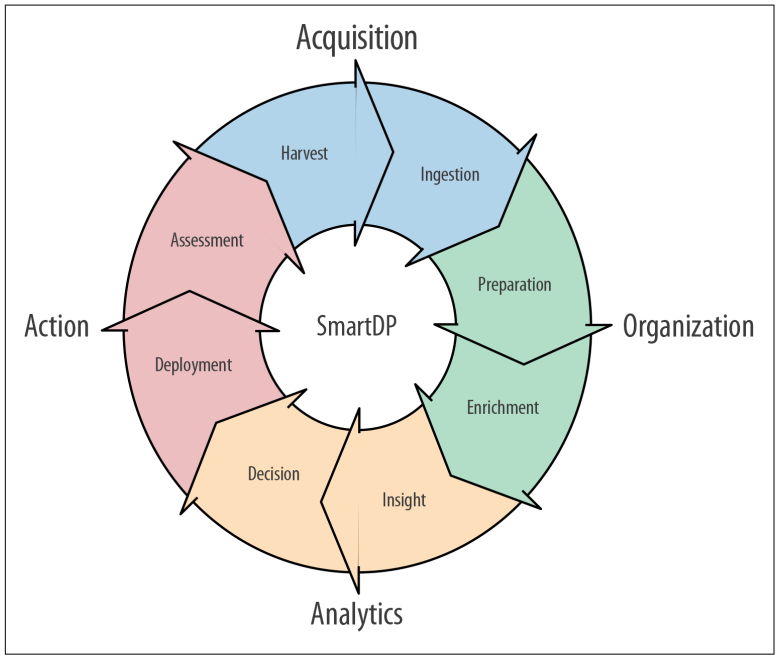

Figure 4-1. Closed loop of data processing (figure courtesy of Wenfeng Xiao)

Data Acquisition

Data acquisition focuses on generated data and captures data into the system for processing. It is divided into two stages—data harvest and data ingestion.

Different data application contexts have different demands for the latency of the data acquisition process. There are three main modes:

Real time
: Data should be processed in a real-time manner without any time delay. Normally, there would be a demand for real-time processing in trading-related contexts. For example:

- For online trade fraud prevention, the data of trading parties should be dealt with by an anti-fraud model at the fastest possible speed, so as to judge if there is any fraud, and promptly report any deviant behavior to the authorities.
- The commodities of an ecommerce website should be recommended in a real-time manner according to the historical data of clients and the current web page browsing behavior.
- Computer manufacturers should, according to their sales conditions, make a real-time adjustment of inventories, production plans, and parts supply orders.
- The manufacturing industry should, based on sensor data, make a real-time judgment of production line risks, promptly conduct troubleshooting, and guarantee the production.

Micro batch
: Data should be processed by the minute in a periodic manner. It is not necessary that data is processed in a real-time manner. Some delay is allowed. For example, the effect of an advertisement should be monitored every five minutes so as to determine a future release strategy. It is thus required that data should be processed in a centralized manner every five minutes in aggregate.

Mega batch
: Data should be processed periodically with a time span of several hours, without a high volume of data ingested in real time and a long delay in processing. For example, some web pages are not frequently updated and web page content may be crawled and updated once every day.

Streaming data is not necessarily acquired in a real-time manner. It may also be acquired in batches, depending on application context. For example, the click event stream of a mobile app is uploaded in a continuous way. However, if we only wish to count the added or retained stream in the current day, we only need to incorporate all click-stream blogs in that day in a document and upload them to the system by means of a mega batch for analytics.

Data harvest

Data harvest refers to a process by which a source generates data. It relates to what data is acquired. For instance, the primary SDK of the iOS platform or an offline WIFI probe harvests data through data sensing units.

Normally, data is acquired from two types of data sources:

Stream
: Streaming data is continuously generated, without a boundary. Common streams include video streams, click event streams on web pages, mobile phone sensor data streams, and so on.

Batch
: Batch data is generated in a periodic manner at a certain time interval, with a boundary. Common batch data includes server blog files, video files, and so on.

Data ingestion

Data ingestion refers to a process by which the data acquired from data sources is brought into your system, so the system can start acting upon it. It concerns how to acquire data.

Data ingestion typically involves three operations, namely discover, connect, and sync. Generally, no revision of any form is made to numeric values to avoid information loss.

Discover refers to a process by which accessible data sources are searched in the corporate environment. Active scanning, connection, and metadata ingestion help to develop the automation of the process and reduce the workload of data ingestion.

Connect refers to a process by which the data sources that are confirmed to exist are connected. Once connected, the system may directly access data from a data source. For example, building a connection to a MySQL database actually involves configuring the connecting strings of the data source, including IP address, username and password, database name, and so on.

Sync refers to a process by which data is copied to a controllable system. Sync is not always necessary upon the completion of connection. For example, in an environment which requires highly sensitive data security, only connection is allowed for certain data sources. Copying is not allowed for that data.

Data Organization

Data organization refers to a process to make data more available through various operations. It is divided into two stages, namely data preparation and data enrichment.

Data preparation

Data preparation refers to a process by which data quality is improved using tools. In general cases, data integrity, timeliness, accuracy, and consistency are regarded as indicators for improvement so as to make preparations for further analytics.

Common data preparation operations include:

- Supplement and update of metadata
- Building and presentation of data catalogs
- Data munging, such as replacement, duplicate removal, partition, and combination
- Data correlation
- Checking of consistency in terms of format and meaning
- Application data security strategy

Data enrichment

In contrast to data preparation, data enrichment shows more preference to contexts. It can be understood as a data preparation process at a higher level based on context.

Common data enrichment operations include:

Data labels
: Labels are highly contextual. They may have different meanings in different contexts, so they should be discussed in a specific context. For example, gender labels have different meanings in contexts such as ecommerce, fundamental demography, and social networking.

Data modeling
: This targets the algorithm models of a business—for example, a graph model built in order to screen the age group of econnoisseurs in the internet finance field.

Data Analytics

Data analytics refers to a process by which data is searched, explored, or displayed in a visualized manner based on specific problems so as to form insight and finally make decisions. Data analytics represents a key step from data conversion to action and is also the most complicated part of data engineering.

Data analytics is usually completed by data analysts with specialized knowledge. Figure 4-2 highlights some key aspects of analytics that are utilized to obtain policymaking support.

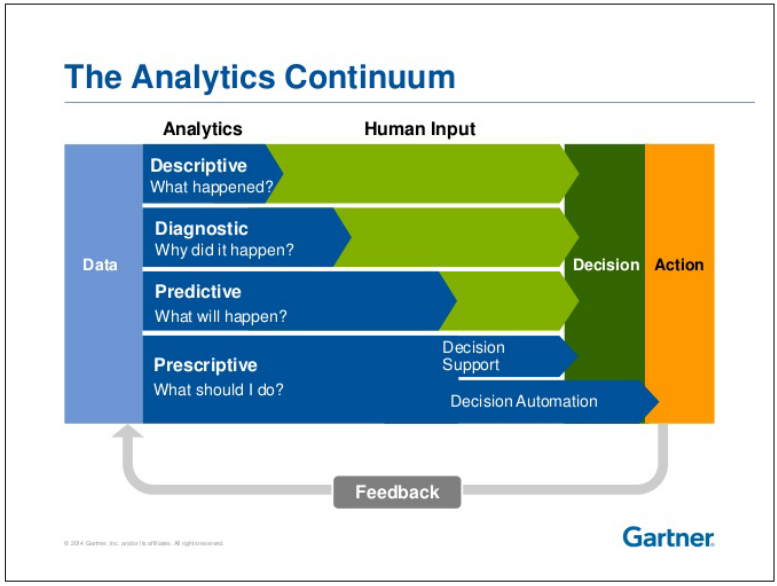

Figure 4-2. Data analytics maturity model (source: Gartner)

Each process from insight to decision is based on the results of this analysis. Nevertheless, each level of analytics means greater challenges than that of the previous one. If the system fails to complete such analytics on its own, the intervention of human wisdom is required. A data analytics system should continuously learn from human wisdom and enrich its data dimensions and AI so as to solve these problems and reduce the cost of human involvement to the largest extent. For example, in the currently popular internet finance field, big data and AI algorithms can be used to evaluate user credit quickly and determine the limit of a personal loan, almost without

human intervention. And the cost of such solutions is far lower than that of traditional banks.

Data analytics is divided into two stages, namely data insight and data decisions.

Data insight

Data insight refers to a process by which data is understood through data analytics. Data insights are usually presented in the form of documents, figures, charts, or other visualizations.

Data insight can be divided into the following types depending on the time delay from data ingestion to data insight:

Real time
: Applicable to the contexts where data insight needs to be obtained in a real-time manner. Server system monitoring is one example of simple contexts. An alarm and response plan should be immediately triggered when key indicators (including magnetic disk and network) exceed the designated threshold. In complicated contexts such as P2P fraud prevention, a judgment should be made if there is any possibility of fraud according to contextual data (the borrower's data and characteristics) and third-party data (the borrower's credit data). Also, an alarm should be trigged based on such judgment.

Interactive
: Applicable to the context where the insight needs to be obtained in an interactive manner. For example, a business expert cannot get an answer in one query when studying the reason for the recent fall in the sales volume for a particular product. A clue needs to be obtained through continuous query, thus determining the target for the next query. The response speed of the query should be in an almost real-time manner, as required by interactive insight.

Batch
: Applicable to the context where the insight should be completed once every time interval. For example, there are no real-time requirements for behavior statistics of mobile app users (including add, daily active, retain) in general cases.

The depth and completeness of data insight results greatly affects the quality of decisions.

Data decisions

A decision is a process by which an action plan is formulated based on the result of data insight. In the case of sufficient and deep data insight, it is much easier to make a decision.

Action

An action is a process by which the decision generated in the analytics stage is put into use and the effect is assessed. It includes two stages, namely deployment and assessment.

Deployment

Deployment is a process by which action strategies are implemented. Simple deployment includes presenting the visualized result or reaching users during the marketing process. However, the common deployment is more complicated. Usually, it relates to shifting the data strategy from natural accumulation to active acquisition. The data acquisition stage involves deployment through construction, including offline construction of IoT devices (especially beacon devices, including iBeacon and Eddystone) and WiFi probe devices as well as improvement of business operation flow so as to obtain specific data points (such as capture of Shake, QR code scanning, and WiFi connection events).

Assessment

Assessment is a process by which the action result is measured; it aims to provide a basis for optimization of all data engineering.

In practice, although the problems of the action result appear to be derived from the decision, they are more a reflection on data quality. Data quality may relate to all the stages of data engineering, including acquisition, harvest, preparation, enrichment, insight, decision, and action. Thus, it is necessary to track the processing procedures of each link, which can help to locate the root causes for problems.

Sometimes, for the purpose of justice and objectivity, enterprises may employ third-party service providers to make an assessment; in this situation, all participants in the action should reach a consensus on the assessment criteria. For example, an app advertiser finds that users of a particular region have a large potential value through analytics, and thus hope to advertise in a targeted way in this region. In the marketing campaign, the app advertiser employs the third-party

monitoring service to follow up on the marketing effect. The results indicate that quite a lot of activated users are not within the region. Does this mean an erroneous action was taken in the release channel? It is discovered through further analysis that the app advertiser, channel, and the third-party monitoring service provider are not consistent in the standards for judging the position of the audience. In the mobile field, due to the complex network environment and mobile phone structure (applications and sensors may be affected), enterprises should pay particular attention to the adjustment of positions, especially when looking at deviations in assessments.

Roles of the Data Engineering Team

Different from traditional enterprises, data-driven enterprises should have their own specialized data engineering teams (see Figure 4-3).

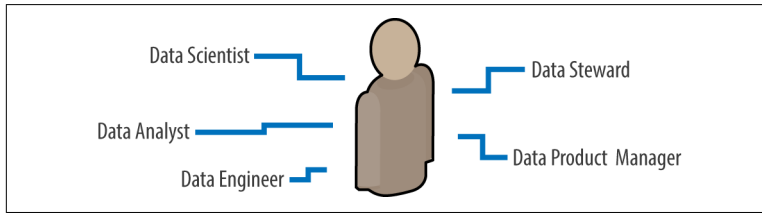

Figure 4-3. Roles of the data engineering team (figure courtesy of Wenfeng Xiao)

A data engineering team involves multiple data-related roles. All members of the team realize data value through effective organizational collaboration.

Data stewards
: As the core of the basic architecture of a data engineering team, data stewards conduct design and technical planning for the overall architecture of the base platform for smart data, ensuring the satisfaction of the entire system's requirements for continuous improvement of data storage and computing when an enterprise transforms itself toward intelligence and continuous development.

Data engineers
: Data engineers are a core technical team that supports data platform building, operation, and maintenance, as well as data pro-

cessing and mining, ensuring the stable operation of the smart data platform and high-quality data.

Data analysts
: Data analysts are the core staff responsible for technology, data, and business; they mine and feed back problems based on analysis of historical data and provide decision-making support for problem solving and continuous optimization in terms of business development.

Data scientists
: Data scientists further mine the relationship between data endogeneity and exogeneity and support data analysts' in-depth analysis based on algorithms and models; they also create models based on data and provide decisions for the future development of an enterprise.

Data product managers
: Data product managers analyze and mine user demands, create visualized data presentations on different functions of an enterprise, such as management, sales, data analysis, and development, and support their decision making, operation and analysis, representing the procedure- and information-oriented commercial value of data.

An enterprise may maintain a unified data engineering team across its organization depending on the complexity of the data and business, or it may establish a separate data engineering team for each business line, or use both styles. For example, the Growth Team directly led by the Facebook CEO is divided into two sub-teams, namely "data analysis" and "data infrastructure." All data of the enterprise should be acquired, organized, analyzed, and acted on so as to facilitate continuous business optimization. Similarly, at Airbnb, the Department of Data Fundamentals manages all corporate data and designs and maintains a unified data acquisition, labeling, unification, and modeling platform. Data scientists are distributed within each business line and analyzed business value based on a unified data platform.

Data steward

A data steward is responsible for planning and managing the data assets of an enterprise, including data purchases, utilization, and

maintenance, so as to provide stable, easily accessible, and high-quality data.

A data steward should have the following capabilities:

- A deep understanding of the data managed and understanding of such data beyond all other personnel in the enterprise
- An understanding the correlation between data and business flow, such as how to generate and use data in the business flow
- Ability to guarantee the stability and availability of data
- Ability to formulate operating specifications and security strategies concerning data

A data steward should understand business and the correlation between data and business, which can help the data steward reasonably plan data—for example, whether more data should be purchased so as to increase coverage, whether data quality should be optimized so as to increase the matching rate, and whether the data access strategy should be adjusted so as to satisfy compliance requirements.

Data engineer

A data engineer is responsible for the architecture and the technical platform and tools needed in data engineering, including data connectors, the data storage and computing engine, data visualization, the workflow engine, and so on. A data engineer should ensure data is processed in a stable and reliable way and provide support for the smooth operation of the work of the data steward, data scientist, and data analyst.

The capabilities of a data engineer should include, but are not limited to the following aspects:

- Programming languages, including Java, Scala, and Python
- Storage techniques, including column-oriented storage, row-oriented storage, KV, document, filesystem, and graph
- Computing techniques, including stream, batch, ad-hoc analysis, search, pre-convergence, and graph calculation
- Data acquisition techniques, including ETL, Flume, and Sqoop

- Data visualization techniques, including D3.js, Gephi, and Tableau

Generally, data engineers come from the existing software engineering team but should have capabilities relating to data scalability.

Data scientist

Some enterprises would classify data scientists as data analysts, as they undertake similar tasks (i.e., acquiring insight from data to guide decisions).

In fact, the roles do not require completely identical skills. Data scientists should cross even higher thresholds and should be able to deal with more complex data contexts. A data scientist should have a deep background in computer science, statistics, mathematics, and software engineering as well as industry knowledge and should have the capacity to undertake algorithm research (such as algorithm optimization or new algorithm modeling). Thus, they are able to solve some more complex data issues, such as how to optimize websites to increase user retention rate or how to promote game apps so as to better realize users' life cycle value.

If we say data analysts show a preference for summaries and analytics (descriptive and diagnostic analytics), data scientists highlight future strategic analytics (predictive analytics and independent decision analytics). In order to continuously create profits for enterprises, data scientists should have a deep understanding of business.

The capabilities owned by a data scientist should include, but are not limited by, the following aspects:

- Programming languages, including Java, Scala, and Python
- Storage techniques, including column-oriented storage, row-oriented storage, KV, document, filesystem, and graph
- Computing techniques, including stream, batch, ad-hoc analysis, search, pre-convergence, and graph calculation
- Data acquisition techniques, including ETL, Flume, and Sqoop
- Machine learning techniques, including TensorFlow, Petuum, and OpenMPI

- Traditional data science tools, including SPSS, MATLAB, and SAS

In terms of engineering skills, data scientists and data engineers have a similar breadth; data engineers, however, have a greater depth. Though high-quality engineering is not strictly required in these contexts, proper engineering skills can help improve the efficiency of some data exploration tests.

As a consequence, the costs paid for obtaining engineering skills are even less than those of communication and collaboration with data engineering teams.

Additionally, data scientists should have a deep understanding of machine learning and the data science platform. However, this does not mean data engineers do not need to understand algorithms as data scientists in some enterprises are responsible for algorithm design while data engineers are responsible for realization of algorithms.

Data analyst

Data analysts are responsible for exploring data based on data platforms, tools, and algorithm models and gaining business insight so as to satisfy the demands of business users. They should not only have an understanding of data but also master the specialized knowledge regarding business such as accounting, financial risk control, weather, and game app operation.

To some extent, data analysts may be regarded as data scientists at the primary level but do not need to have a solid mathematical foundation and algorithm research skills. Nevertheless, it is necessary for them to master Excel, SQL, basic statistics, and statistical tools, as well as data visualization.

A data analyst should have the following core capabilities:

Programming
: A general understanding of a programming language, such as Python and Java, or a database language can effectively help data analysts to process data in a scaled and characteristic manner and improve the efficiency of analytics.

Statistics
: Statistics can enable analysts to interpret both raw data and business problems.

Machine learning
: Regular machine learning algorithms can help data analysts to mine the potential interaction among groups in the dataset and improve the depth of their business analysis and breadth of perspectives.

Data munging
: In view of the many nonendogenetic controllable factors that may cause data deviation, such as system stability and manual operation, data analysts are required to mung data in a professional manner and guarantee the data quality of data analysis.

Data visualization
: Data visualization supports data interpretability for management decision making and salespeople on the one hand and enables data analysts to conduct a second mining of business problems from the perspective of charts on the other hand.

Generally, data analysts are experienced data experts in teams. They should have a strong curiosity about data and stay aware of the continuously emerging techniques and best practices.

Data analysts should also be skillful at communication. They should not only be able to satisfy the demand for exchanging business with business users but also maintain a smooth collaboration with the other roles of data teams so as to obtain resource support.

Data product manager

A data product manager focuses on data value. Data product managers should understand business, data, and their correlation. Thus, it is required that a data product manager have the capabilities of both product manager and data analyst.

Data product managers are not data analysts but should collaborate with data analysts and make data engineering product-oriented. Data analysts focus more on projects as well as the efficiency and speed of insight. They use the most convenient and agile tools to deal with business problems as soon as possible. Data product managers focus more on products and the stability and reliability of insight. They should select validated techniques and consider data

security. Data product managers should also consider business demand more completely and develop data products that can be recognized in abnormal environments.

Data Science

As required by the smart data era, data science spans across computer science, statistics, mathematics, software engineering, industry knowledge, and other fields. It studies how to analyze data and to gain insight.

With the emergence of big data, smart enterprises must deal with a greater data scale and more complex data types on smart data platforms through data science. Some traditional fields and data science share similar concepts, including advanced analytics, data mining, and predictive analytics.

Data science continues some ideas of statistics, for example, statistical search, comparison, clustering, classification, and other analytics and summarization of a lot of data. Its conclusions are correlation rather than a necessary cause–effect relationship. Although data science relies heavily on computation, it is not based on a known mathematical model, which is different from computer simulation. Instead, it replaces cause–effect relationship and rigorous theories and models with a lot of data correlation and acquires new "knowledge" based on such correlation.

CHAPTER 5
SmartDP Solutions

Normally, a complete set of SmartDP solutions involves data, platform products, data applications, and consulting/services.

Data Market

Data requires flow, interaction, and integration to bring its largest value into play. The data exchange and trade market enables data suppliers to upload, introduce, publicize, and transfer the transmission of data and let purchasers try out, inspect, and acquire data in scale, representing the middleman and guarantor in the trade.

In addition to necessary measuring and billing, more crucial characteristics of the data market include solving the problems of data conversion, conforming to laws and regulations, fraud prevention, standard unification, quality verification, data convergence, and so on.

The key functional points of the data market include:

Conforming to laws and regulations

- Incorporating a checking mechanism to avoid the personal identifiable information–type (PII) data and easing privacy issues through asymmetric cryptography
- Providing a guarantee and a conversion platform to assist both buyer and seller in obtaining data results under the status of "available but invisible"

- Validly processing data through ID conversion

Fraud prevention
: Increasing anti-fraud rules and avoiding the data fraud of some suppliers

Standard unification
: Providing data input and standards for input interfaces as well as industry-based and type-based standards for business data

Quality verification
: Verifying data quality through cross verification, business feedback, sampling, spot inspections, etc.

Data convergence
: Converging and converting scattered small data sources and increasing the availability of data.

Platform Products

SmartDP products should be able to support data management, data engineering, and data science. The platform should not only satisfy the requirements for data management, but also complete data acquisition, organization, analytics, and actions and support the building of data science algorithms and models. In terms of ecology, the platform is capable of generating smart data applications and providing support for the data market through data processing and production capabilities.

Data Management

In terms of data management, SmartDP products should include the following three elements:

Multisource gathering
: Self-owned data, self-owned data on third-party platforms, third-party data

Quality enrichment
: Mainly monitoring and capacity enhancement of data quality

Strategy control
: Including management and control of data assets (such as safety and validity, and access security and authority)

Data Engineering

In terms of data engineering, SmartDP products should include the following elements:

Infrastructure
: In order to respond to the flexible data engineering environment, the platform should have flexible deployment capabilities and support deployment modes including private, cloudization, sandbox, etc.

Introduction of techniques
: Leading data techniques should be introduced to increase the storing and computing efficiency of multivariate data, thus improving the efficiency of data engineering.

Wisdom of the field
: Enterprises should change their thinking mode from data operation to operating data. It is thus required that human experience and the wisdom of the industry be incorporated into data engineering.

Data Science

In terms of data science, SmartDP products should include the following elements:

Training mode
: As the initial link of data science, the platform should be able to construct, verify, and adjust models based on the analysis results of business demand.

Precision of verification
: As the verification link of data science, the platform should be able to confirm the precision of the model based on prior data and human experience.

Application of execution
: As the application link of data science, the platform should be able to realize the execution, operation, and maintenance of algorithm models.

Data Applications

Data is applied in the smart data age are utilized to satisfy the logical packaging of business needs with SmartDP's capabilities (data management, data science, and data engineering) in order to solve specific business issues and realize business values.

Data applications are usually undeveloped by data product managers who provide services to business users. Therefore, they need to consider user experience, including the simplicity of operation process and the clarity of visual delivery.

In practice, data applications may be derived either from enterprises or from third parties. Third-party data application providers may have proven capacity in some vertical fields, such as financial risk control, customer value prediction algorithms, and context awareness, which could supplement the experience of enterprises in these fields and avoid reinvention. However, when insufficient third-party data applications exist in the market to solve practical business context problems, enterprises must realize customized data applications in a targeted way through independent development or subcontracting.

Data applications can support each other to help realize the reusability of data and functions.

Consulting and Services

Generally speaking, SmartDP solutions may integrate into users' businesses and marketing departments by means of consulting and convert users' problems into those that can be solved by data and data applications, in order ultimately to provide support for users' decisions. SmartDP aims to provide the basis and methods for informed decision making (including objective and subjective decisions). Thus, consulting may be the primary usage of SmartDP.

Relationship of SmartDP with Existing Data Platforms

The relationship between SmartDP and existing data platforms (including data warehouses) should be seamless. No new platform is required to replace the existing one. In fact, SmartDP constitutes a supplement to the existing data platform and could enable the exist-

ing platform to operate in a better way. The following are some ways this is possible:

- SmartDP and the existing data platforms can be data sources of each other. The sorted data sources on SmartDP can be opened in a safe and convenient way via an interface.
- SmartDP is able to deal with massive data of multiple dimensions and types while the current data platforms of enterprises are generally good at dealing with smaller, more structural data. The two can supplement each other.
- SmartDP can provide the existing data platforms of enterprises with data science capabilities, such as predictive analysis and data mining.
- SmartDP has lower-cost storage and computing capacities and strong and flexible scalability, which can reduce the cost pressures of the existing data platform in terms of storage and computing.

Commercial Models of SmartDP

As mentioned earlier, SmartDP is a platform for exploring the commercial value of data based on smart data applications. The commercial value and smart data applications represent two key points that may contribute to a reasonable commercial model. The possible commercial models for SmartDP may include:

Data exchange and transaction
: More fragmented data suppliers may sell their data in the data market, which is a more advanced new commercial model. Both data exchange and transaction platforms should play multiple key roles including transaction guarantee, privacy protection, security and legality, measurement and billing, quality verification, fraud prevention, convergence improvement, and so on, and charge a commission during the transaction.

SAAS
: Tools of smart data applications and their business value are regarded as a software service to be sold. A service fee is usually charged regularly.

Software and system sales
: Software and systems are sold to enterprises as technical tools. There is a one-time charge for delivery and annual maintenance.

Profit sharing
: Profit sharing is actually the most confident means in commercial model. It collects a certain percentage of the incremental revenues earned by enterprises. For smart data applications, suppliers should bear data source–related costs and application design costs on their own. It is required that specific businesses be further developed to bring an increase in revenue. Also, value and revenues are subject to quantitative assessment. In good operating conditions, this model would mean the largest revenue for both buyer and seller.

CHAPTER 6

SmartDP Reference Architecture

In terms of architecture, SmartDP should be able to support data management, data science, and data engineering, and enable data engineering teams to effectively collaborate with one another (see Figure 6-1).

Functionally, SmartDP is divided into five layers, as shown in Figure 6-1—the data layer, data access layer, infrastructure layer, data application layer, and operation management layer. Let's look at each layer in detail.

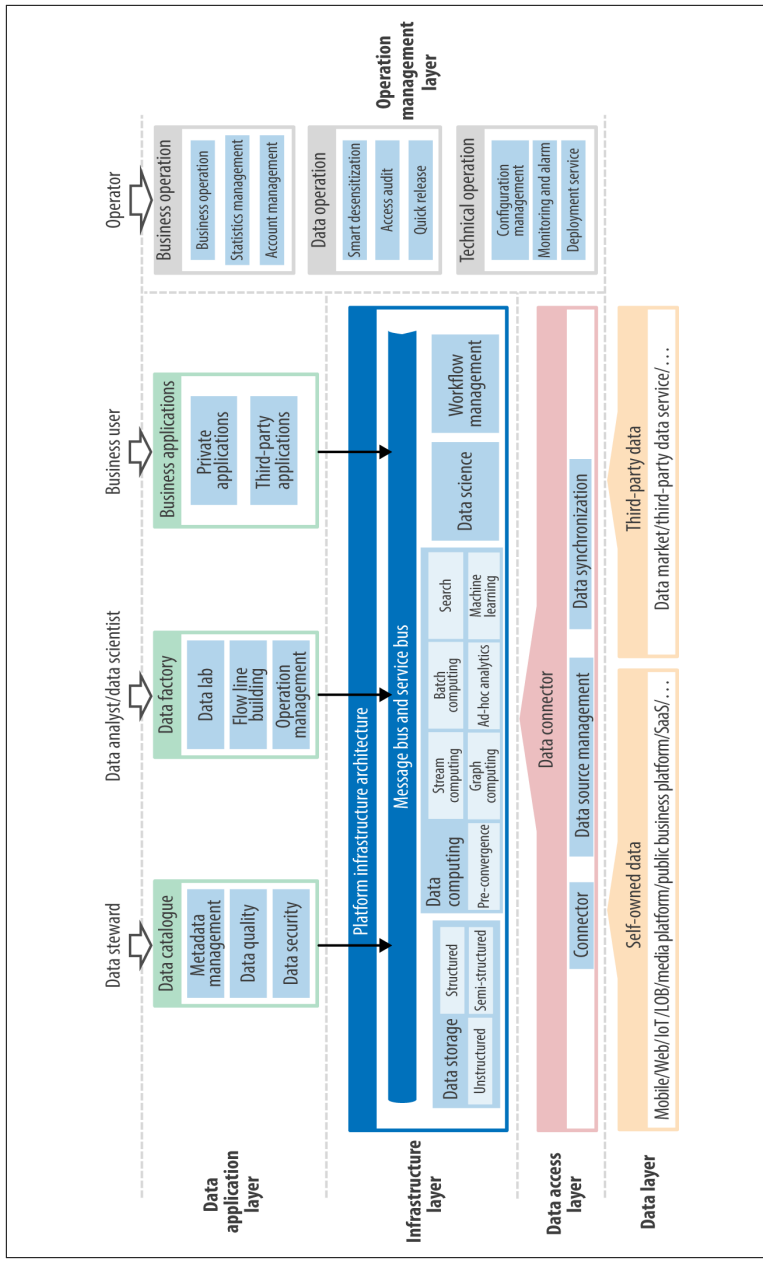

Figure 6-1. SmartDP reference architecture (figure courtesy of Wenfeng Xiao)

Data Layer

The data layer may be divided into self-owned data and third-party data, according to sources.

Self-Owned Data

Self-owned data refers to data that is owned by an enterprise and that can be completely controlled and managed by the enterprise, which is stored in the enterprise or any external platform of a third party.

The self-owned data stored in an enterprise generally includes:

- Data from mobile smart devices, including user interaction data of native smart phone applications (including Android, iOS, WinPhone) and of vertical application frameworks (including Cocos2D, Unity3D, APICloud, Cordova).
- Data from web pages, including user click events of PC or mobile web pages (such as JavaScript) and web page information crawled from web pages
- Data from IoT devices, including beacon devices (such as iBeacon and Eddystone) and user visit data captured by WiFi probes
- Data from the existing systems of the enterprise, including CRM systems, ERP systems, and data warehouses

The self-owned data stored outside an enterprise includes:

- Data on public media platforms, including the WeChat public platform, Microblog platform, Twitter, and Facebook
- Data on public business platforms, including the goods trading data of Tmall, media data from advertising alliances, and coupon consumption data of the O2O platform
- Data stored on public SaaS platforms, including data from advertising monitoring platforms, user analysis data from mobile phone app analysis platforms, and customer data from customer management systems

For the harvest of first-party data, enterprises should have standard data harvest techniques so as to stay compatible with data harvest

methods in mainstream environments and platforms. The techniques include but are not limited to:

- Mainstream smart phone platform support, including Android, iOS, and WinPhone
- Third-party platform support, including Cocos2D, Unity3D, APICloud, and Cordova
- Web page support, including JavaScript and web page crawlers
- Support from offline beacon devices (including iBeacon and Eddystone) and WiFi probes.
- Context-awareness data support from smart devices, including human gestures (such as walking, running, driving, riding, handholding)

Third-Party Data

Third-party data is the data not owned by an enterprise and provided by a third-party data provider.

Normally, the third-party data includes:

- Data purchased by an enterprise in the data market, such as mobile phone label data of a third-party DMP
- Data services provided by a third-party data provider, such as traffic data ingested via government data open interface

Data Access Layer

The data access layer ingests the data harvested into the system for subsequent operation. Thus, in order to gain access to more data sources, the data access layer should support as many data connection types as possible, which include but are not limited to the following aspects:

- Connection support for frontend event harvest sources, such as WiFi probe data, mobile phone interaction events, and web page click events
- Connection support for various public media platforms, public business platforms, and public SaaS platforms

- Connection support for RDBMSs
- Connection support for data warehouses and OLAP (MDX)
- Connection support for local filesystems
- Connection support for common big data systems (including Hadoop/HBase)
- Connection support for NoSQL databases (such as Cassandra and MongoDB)
- Connection support for traditional analysis platforms, such as SAP/Siebel
- Connection support from buffer services (such as Redis)

Infrastructure Layer

The infrastructure layer provides the basic technical capacity for data applications at upper layers, including data storage, data computing, data science, workflow management, the message bus, and the service bus.

Data Storage

Data storage capacity supports long-term data storage.

Data storage can be divided into the following types according to the type of data stored:

Structured
: It is required that a predefined strict data model (schema) or a predefined organization mode be established for storage. Normally, the structured storage includes row-oriented storage (including traditional RDBMSs, and MySQL), column-oriented storage (including column-oriented databases, such as HBase, and Vertica) and graph storage (including graph databases, such as GraphSQL).

Semi-structured
: This has no strict definition of a data model but a certain format, which is freely scalable, such as JSON and XML.

Unstructured
: There is no strict definition of a data model. And there is no way to pre-organize the storage type. Unstructured storage

includes scalable or freely organized files, including server blogs, emails, compressed files and videos.

Data storage can be divided into the following types in terms of the form of stored content:

- Filesystem, such as HDFS and Alluxio, applicable to the storage of large-scale datasets
- Graph storage, such as Neo4J, applicable to data structures with multilevel interaction
- Column-oriented storage, such as HBase and Parquet, applicable to batch data processing and real-time query
- Row-oriented storage, such as MySQL, applicable to random query
- KV storage, such as Redis and Aerospike, applicable to business data with fewer data relations
- File storage, such as MongoDB, applicable to the storage and query of document formats

Data Computing

Data computing capacity supports all operations of data organization and analytics.

Data computing can be divided into the following types based on computing type:

- Stream computing: Storm, Spark
- Batch computing: Hive, Spark
- Ad-hoc analytics: Presto
- Search: Elastic Search
- Pre-convergence computing: Druid, TalkingData, AtomCube
- Graph computing: GraphSQL
- Machine learning platforms: TensorFlow, Petuum, OpenMPI, Spark

Data Science

Data science capacity provides various algorithms and models.

According to the application contexts of algorithms and models, data science can be divided into the following algorithms:

- Data standardization algorithms: TF-IDF weighted statistics
- Time-frequency transformation algorithms: fast Fourier transformation
- Dimension reduction algorithms: principal components analysis
- Clustering algorithms: hierarchical clustering, power iteration clustering, streaming *k*-means clustering, Gaussian mixture model, latent Dirichlet allocation, and DB scan
- Classification algorithms: naive Bayes classification, *k*-nearest neighbors, support vector machines, batch logistic regression, simple linear regression, random decision trees, Lasso regression, decision trees, and logistic regression
- Recommendation algorithms: collaborative filtering
- Interaction analysis algorithms: a priori data mining, FP-Growth and cSPADE
- Algorithm models owned by other enterprises

Workflow Management

Workflow management is used to manage data-related work and flow processes. It arranges various automatic or manual tasks necessary for data engineering and ensures their completion and implementation. Workflow management relates to task security, error handling, dependency arbitration, and other coordination tasks.

The common tools for workflow management include Azkaban and Oozie.

Message and Service Bus

The message bus aims to decouple message generators and consumers and the service bus aims to decouple service providers and callers. The message and service buses can provide the system with a more flexible architecture and much higher scalability.

The common message buses includes Kafka, and the common service buses include such microservice governance frameworks as Eureka and Dubbo.

Data Application Layer

Data applications are the commercialization of data engineering that satisfies particular demands. They may run normally based on the data convergence of SmartDP and technical capacity of the platform.

SmartDP includes both system applications similar to data catalogs and data factories, as well as business applications of various types.

Data Catalog

Equivalent to a search engine for data, the data catalog can facilitate users to search, understand, and use data by organizing the relevant information of data.

Generally, the data catalog is maintained by the data steward and may be used by all other roles of the data engineering team.

Normally, the data catalog undertakes the tasks as data management and includes the following core functions:

Metadata management
: Management of data lineage, data summary, and data format

Data quality management
: Management of record rearrangement, unit error correction, and data correlation

Data security management
: Management of audit, hiding, and tokenization and access control

Data Factory

A data factory is used to build production flow lines for data engineering and ensure these lines operate normally.

Generally, a data factory is maintained by data engineers and may be used by all other roles of the data engineering team.

Normally, a data factory includes the following core functions:

Data lab
: For exploring data engineering, including explorative data analysis, building and verification of new algorithm models, and trying new data processing procedures

Flow line building
: For assembly data processing, including visualized task scheduling and interactive exploration

Operation management
: For managing all flow lines, including smart task deployment, monitoring and abnormal alarms, and heterogeneous computing

Business Applications

Business applications provide services to business departments.

In general, business applications are used by business users and may be developed either by the enterprise or by a third party.

Common business applications include marketing release, mobile analysis, advertisement monitoring, offline visitor flow analysis, competition analysis, financial risk control, and identity verification.

Operation Management Layer

The operation management layer is able to support the business, data, and technical operations of SmartDP.

Business Operations

Business operations can help operators manage the SmartDP account as well as the measurement and billing of services provided externally.

Data Operations

Data operations can enable operators to manage compliance, anonymization, and access control rules as well as the audit strategies of the data services they provide, and improve the efficiency of data service releases.

Technical Operations

Technical operations may help operators manage server configuration and monitor alarm strategies at the core of SmartDP as well as maintain the deployment of platform resources.

CHAPTER 7

Case Studies

SmartDP Drives Growth in Banks

Smart data is important to the financial industry, a highly information-oriented industry. If we compare the finance industry to a car, the information system is the engine of the car and data the fuel. By taking advantage of smart data, the financial industry will witness faster growth, reduced costs, and more first-mover advantages.

Financial enterprises have experienced business engineering and technical engineering in their development process. Business engineering refers to a process by which business models are searched, mainly for the establishment of a business process. Technical engineering refers to a process by which the business process becomes tool- and system-oriented during the business engineering implementation. Both business engineering and technical engineering solve the issue of modern and standardized production.

With the aid of TalkingData, a bank built its big data application using the SmartDP model. Unstructured data was acquired and analyzed and then labeled according to data application contexts. A machine learning model was established based on trade data and behavior data to satisfy the demands of banks for large-scale data mining and analytics. In contrast to the traditional BI tools, SmartDP is able to process both structured and unstructured data with a capacity of 100T or above and also user behavior data of hundreds of millions of dimensions. Using the Atom computing engine,

independently developed by TalkingData, a query and computation of millions of pieces of data takes only minutes to complete.

Multisource Data Concentration and ETL

The bank we observed owns plenty of data in terms of personal attributes, assets, credits, and trades but lacks the financial and behavior data of users at other financial institutions. The original information systems such as the CRM system, product system, payment system, bookkeeping system, and channel system gather data in the existing data warehouse. The SmartDP system, according to its business demands, retrieves data from the data warehouse and establishes a table of data from all data sources for data applications. Harvested unstructured data and external data also concentrate in SmartDP for unified data connections and extract, transform, and load (ETL). The SmartDP system helps the bank to integrate internal trading data, customer attribute data, behavior click data, and external data; establish a list of data assets; and manage data as an asset. Massive data harvest and the convergence of banks and data ETL occur here in real time. The SmartDP system also supports real-time data analysis and presentation.

The SmartDP system can accomplish the convergence and connection of more than three types of data source and tries to help engineers with multi-data source integration and munging semi-automatically by machine learning. When an ETL engineer conducts data integration and munging, a machine learning module capable of active learning will monitor and learn the whole process. For example, with the help of machine learning, it can judge whether an 11-digit figure may be a mobile phone number—some fields may represent price and some fields may represent age. Also, it can analyze addresses, names, and transactions. When the machine learning module is trained, it can conduct the heavy work of data munging and processing, thereby replacing ETL engineers. Active learning will enable human engineers to ensure greater ETL accuracy in the near future. This is a real fact rather than a presumption.

Contextual Data Labels and Data Management

SmartDP helps banks integrate multisource data, build corporate data assets, and establish connections among internal structural data, nonstructural data, and external third-party data. The

SmartDP system establishes a management catalog of data assets, labels data according to business demands and contexts, and conducts data management in a unified manner. It builds data labels by visual dragging and realizes data application functions by data analysis and mathematical models. Different from the existing enterprise systems and platforms, the core of SmartDP lies in management and application of data asset. Data asset management involves real-time introduction and connection of multiple data sources, data governance, data quality management, and data processing monitoring.

The contextual labels of SmartDP may be derived from the results of data screening and statistical analysis, the results of logic operations, and even from the results of decision data, such as high-value potential clients, potential groups for financial products, lost high-value clients, home buyers, car buyers, gold buyers, consumption loan borrowers, and even fraud committers and future default clients.

Machine Learning and Data Application

Machine learning can be used to analyze data dimensions in relation to the financial demands of clients. Location data and click/browsing data are particularly important dimensions. Certain seeds of clients are input for learning and behavior data. The groups that are similar to seed clients are found among massive amounts of data. Similar groups have similar social roles and interests. The sales of similar products to similar groups may create a higher conversion rate.

The data captured by smartphone users can identify customers' consumption and interest preferences. By combining a bank's transaction and customer data, the user behavior data can help the bank to search for high-value customers and improve their operating revenues using machine learning. The analysis of nonstructured behavior data may allow the bank to find more data application contexts. The SmartDP system mines high-value potential clients by combing both internal and external data and using a logistic regression model. The model's AUC (Area Under ROC Curve) can reach 0.9 (AUC 0.7 or above means that the model works well and can be adopted in business) and there is a more than 10 times improvement in marketing effect. TDA (topology analysis) and RF (random forest) are used to identify the potential default clients, with AUC reaching 0.8 (both 0.9 and 0.85 are great numbers, meaning that the model will perform very well). Graph knowledge is used to detect

the characteristics of fraud committers, establish an anti-fraud model, and identify 90% or more of the users that commit fraud.

Application Contexts

High-value client mining and marketing

Financial enterprises represent a typical Pareto effect. That is, 20% of their clients contribute 80% of the operating revenues. TalkingData discovered through data analytics that 8% of financing clients for the mobile end of a bank own about 75% of the total assets of the bank. The bank hoped to find more high-value clients for marketing and improvement of financing products sales performance.

With 30,000 high-value clients as seeds and the variables related to high-value clients as input, among millions of mobile devices, TalkingData calculated the devices that are similar to those high-value clients based on the lookalike algorithm of the Atom engine. It gets engaged in marketing by using the Push and SMS functions of the digital marketing tools. In the SmartDP model used to mine high-net-worth clients, TalkingData used data of several dimensions as input variables, including device concentration point, application name, device model, transaction information, and customer information, and searched for high-value potential clients in data with 50 million dimensions.

By this method, the bank sold millions of dollars of financial products within two months. Compared with traditional marketing means, costs were reduced 95% and the bank saw an increase of 15% in high-value clients.

Improving the marketing conversion rate more than 10 times

The bank would invest plenty in marketing costs every year, including red envelope incentives issued to all clients. However, the bank found that there was a low rate of conversion for financial products using the red envelope. Normally, the conversion rate of red envelopes that contributed to sales performance was lower than 0.3%, representing a big waste of red envelopes and marketing time.

Thus, the bank, with clients that had responded to red envelopes, got engaged in machine learning in the existing client information database. Device information that was similar to these seeds was used to locate the potential groups who purchased financing prod-

ucts by red envelope. Through push notifications, the SmartDP system got involved in marketing on these target devices to improve the conversion rate of the red envelope incentive. Through precise marketing with machine learning, the conversion rate of the red envelope incentive saw a more than 10 times improvement.

Before a marketing campaign, TalkingData's SmartDP used machine learning to find 50,000 clients who were candidates to purchase financing products by using an incentive. One week later, the bank found that the conversion rate for financing product purchases was increased from the previous 0.3% to 4.5%, a 15x increase in product marketing conversion rate.

Client loss warning and waking inactive clients

It was found that many clients chose to redeem their financing products from their bank accounts upon their maturity rather than to purchase them again. There was a high client loss rate within a certain period of maturity of the financing products. Some other clients transferred their funds soon after purchasing T+0 (monetary fund) products. The bank wondered where these funds were used and had no way to market.

Using SmartDP, the bank calculated tens of thousands of clients were lost and sent SMSs to them to push their exclusive financing products upon the expiration of their existing products. More than 60% of clients opened the SMS connection and 30% of clients chose to purchase financing products from the bank. As a result, the client loss rate was reduced by 30% and loss of financial products reduced by hundreds of millions of dollars, thus adding interest revenue of millions of dollars for the bank.

SmartDP calculated inactive high net worth clients by using machine learning and considering the active duration of a client's device, financing revenue, amount of assets, and characteristics of high net worth clients. Machine learning algorithms were run once every month to screen the high net worth clients that have become inactive. The marketing push function was used to send exclusive incentive red envelopes to wake inactive clients, activate their transactions, and bring in more assets and transaction charges for the bank. When launching marketing campaigns toward high net worth clients, SmartDP helped the bank to activate more than two billion financing transactions in one month. By using data analytics, the

bank searched for clients that had not yet engaged in transactions or who had become inactive for one year or more and analyzed the behavior characteristics of such clients. It learned about clients' interests by using external data and launched target marketing toward such clients. The bank activates inactive clients with red envelope incentives and game coupons. Over three months of targeted marketing, around 40% of inactive clients were activated, thus bringing handsome revenues to the bank.

The SmartDP system is a perfect combination of data engineering and data application (see Figure 7-1). Functionally, it can help enterprises to introduce and integrate multisource data, process and connect data in a real-time manner, conduct data governance and management, and monitor the quality of data assets and completion of data engineering. In terms of data value application, SmartDP can help enterprises to draw portraits of users, label contextual data according to business demands, and build a closed loop of digital marketing by means of EDM, SMS, and Push. Enterprises may design their marketing campaigns using their marketing management tools and manage these campaigns, including selection, design, sending, and monitoring of digital marketing plans. Also, SmartDP can enable enterprises to learn about the ROI of marketing campaigns and adjust campaigns, their pushing targets, marketing duration, investment budgets, and statistical methods according to marketing feedback and effects.

AI applications can also be integrated in SmartDP. The interaction and transaction data provided by SmartDP can help AI applications optimize their input data and output results and support robo-advisors, automated customer service, and intelligent recommendation engines—all popular applications of AI in banking.

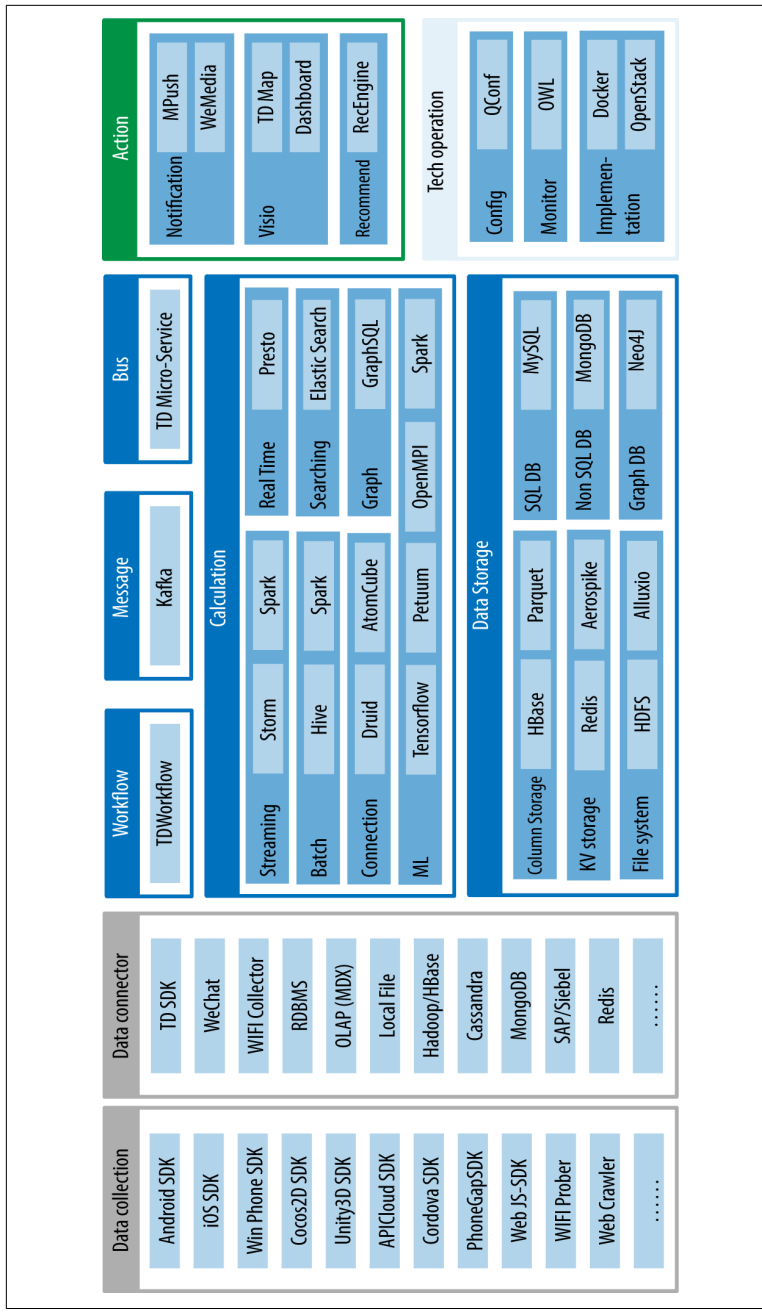

Figure 7-1. Core data engineering technology of SmartDP (figure courtesy of Wenfeng Xiao)

Real Estate Development Groups Integrate Online and Offline Marketing with SmartDP

Real estate developers control massive amounts of home buying information that they have accumulated for years. The effective organization and mining of such information may bring about a new profit space and profit growth point in the big data era. A common consensus has been reached in the real estate industry that big data can generate high value for real-estate developers and agents. In the marketing area especially, no effective methods to harvest and acquire deep insight on client information are available in traditional modes, thus causing failure in the goal of marketing in a targeted manner. SmartDP is used to depict customer preferences and behavior, compare the differences of customer groups in terms of home visits, generate competitive products and transactions, formulate assisted marketing strategies, guide the selection of marketing approaches, and optimize marketing plans.

Common Market Practices and Disadvantages

Due to the limitation of data management and data analytics, a low number of data dimensions, and other factors, a large amount of data accumulated by the real estate industry failed to be fully applied during the traditional homebuying process. During the analysis process, the small sampling method is adopted to sample clients. Basically, the number of samples selected accounts for around 0.04% of all samples, which may lack some key influencing factors and lead to deviation of analysis of customer group characteristics. After sampling, an investigation lasting 7 to 10 days would increase client acquisition costs. The dimensions selected for investigation would be simplistic and the user group would be insufficiently depicted. Due to the fact that static data is used in user group sampling and data ingestion, the analysis results and the actual conditions may conflict with each other. In the final stage of analysis, industry experts would get involved in quantitative description. However, the possible deviation in the previous analysis might heavily influence the judgment of experts. Thus, there is no way to guarantee the objectivity and fairness of the analysis result. During the stage of market reach, leaflets and roadside billboards would be normally adopted for publication, which is not targeted or efficient.

Methodology

With over five years of development and accumulation of data, TalkingData has built its unique methodology (see Figure 7-2).

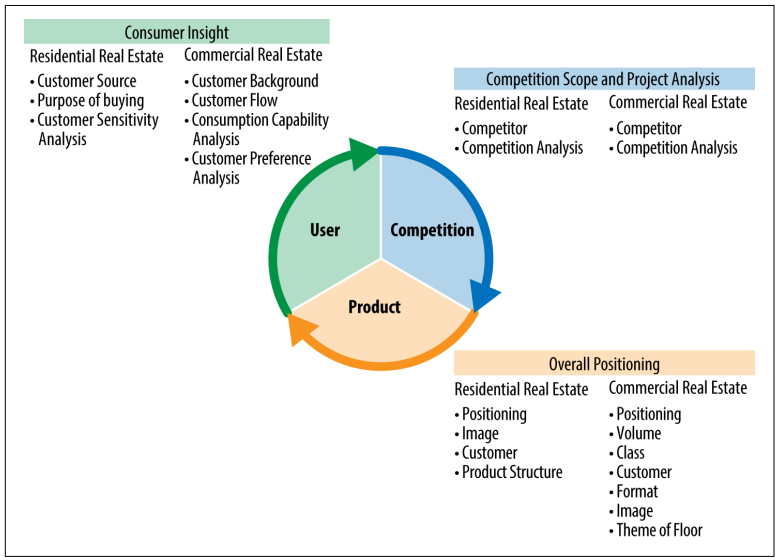

Figure 7-2. TPU methodology (figure courtesy of Yifei Lin)

As shown in Figure 7-2, the TPU (Traffic, Product, and User) methodology highlights the relationship among Channel/Traffic, Product, and User. TalkingData will establish a label system for target groups (for instance, real estate buyers) to profile dimensions such as demographics, wealth, hobbies and interests, brand preference, and real-life locations: establish a network of connections and relations between devices, scenarios, and audiences working with the developer's first-party data dimensions, such as unit models and volume of transactions; and filter for the target groups for future marketing base on these profiles. This label system can be deployed and established in SmartDP, realizing a 360-degree panorama on the target group and driving follow-up marketing based on the aggregation and interconnection of client data and external data.

We can learn about the general situation of a city, understand population traffic, and identify our promotion channels according to urban development and the general trend of population migration. Through analysis and orientation of competitive products, we see a

more informed view of clients (real-estate developers and agents) and more targeted marketing strategies may be formulated.

Description of the Overall Plan

The following SmartDP solutions were formulated by TalkingData after analyzing the demands of target clients (see Figure 7-3).

Figure 7-3 shows a breakdown of the marketing process from the layers of data, platform, operation, and demand. Flexible and efficient DataApps provided by SmartDP (such as a city map DataApp for urban investment strategy, a site visitor flow management DataApp for real estate sales, and a member flow analysis system DataApp for shopping malls) were used to help clients make an accurate and effective analysis. These apps can also help to form a closed loop of industry that covers early investment strategy formulation, product or brand positioning, marketing, and operation, and also a closed loop of marketing (planning, audience selection, campaign execution, and outcome data verification, outcome and revenue analysis).

In terms of data, SmartDP.DMK acquires and organizes various types of human-oriented data (age, gender, hobby preference, brand preference, etc.) necessary for marketing and forms a clean and available dataset.

In terms of platform, SmartDP.DMP integrates and connects the data across the continent from Smart.DMK, analyzes and labels the data, and obtains the basic characteristics of user groups. Also, it manages data and user groups and conducts a dynamic analysis of the data.

The marketing process is broken into three steps. First, it analyzes the characteristics of the visitor group, transaction client base, and competitive product group, and gets potential user groups. Second, it establishes offline populations' application preferences, effectively integrates release channels, and recommends the optimum channel. Third, it reaches clients offline.

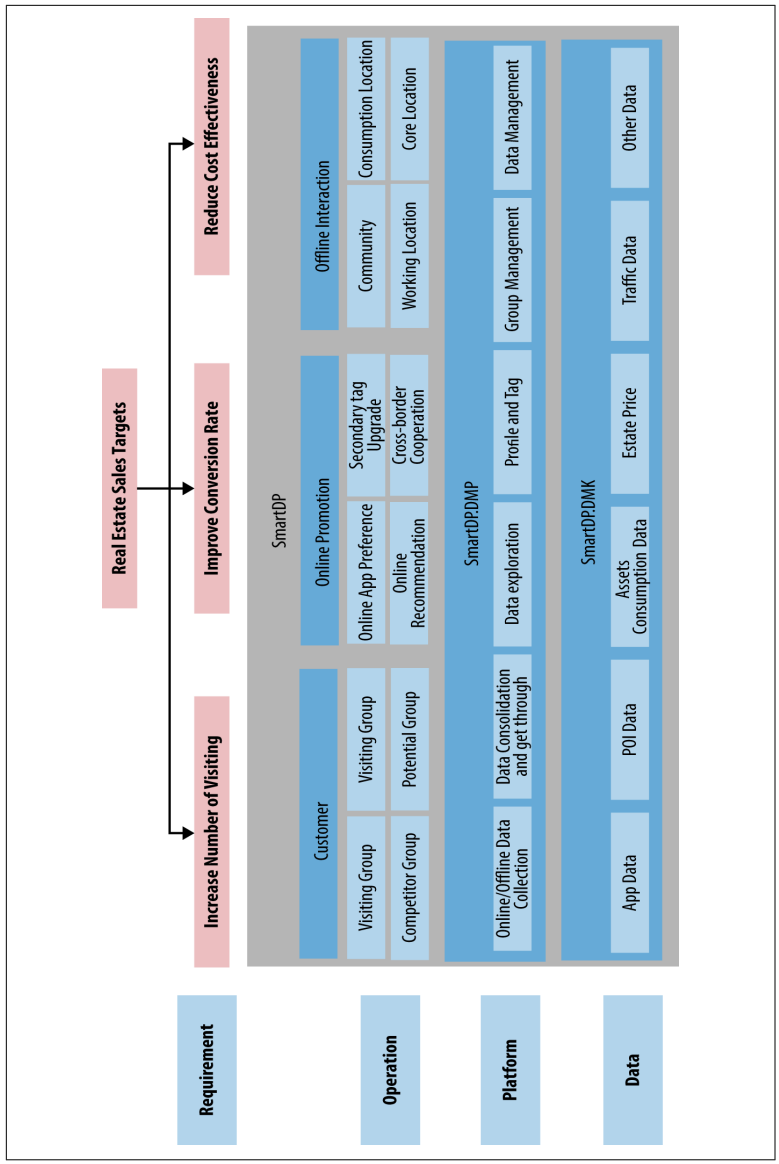

Figure 7-3. Solutions to the marketing of residential houses (figure courtesy of Yifei Lin)

Finally, by implementing SmartDP, TalkingData undertakes a total improvement of the current marketing process, and describes the deep characteristics of the client base in an all-around manner (a client portrait) starting from the accumulation of visitor group data

(acquired by sensors) so as to effectively reach target clients (targeted release). It also effectively monitors feedback (effect validation), and forms a closed loop of marketing. For the closed loop of marketing guided by the "Big Data" concept, data harvest and analysis may be conducted at each key point of the marketing link. Data is used to measure the KPI for each business link. Suggestions for optimization of each indicator are made through data analysis and mining. Also, the subsequent result data is acquired and attributed. These steps can effectively facilitate the core links, including door-to-door visit statistics, conversion rate increase, and cost-effectiveness ratio.

Specifically, during the period of positioning, based on the real estate industry label system from TalkingData, the developer can position potential clients in a targeted manner, describe the characteristics of the client base, and assist in the formulation of the overall marketing strategy. During the stage of accumulating home buyers, we provide aid in management, monitor of the visitor to marketing cases, gain insight on visiting clients, follow up on the analysis of competitive product user groups, and adjust the orientation for customer development. During the stage of continuous marketing optimization, we gain insight on the transaction client base, compare the differences among visiting client bases, adjust online and offline promotion approaches, and optimize marketing plans.

The implementing process and effects of the solutions are detailed next.

Data Harvest and Organization

Harvest of data generated from the sales office of TalkingData's clients is done using WiFi probes that are deployed on site to get accurate information about visiting clients, including their MAC (Media Access Control) addresses, time of arrival, and time of departure.

Data was harvested for a period of three months.

The data acquired by clients (sales data, payment data, contacts, etc.) and the data acquired by TalkingData (public user behavior data from mobile phones) are uploaded in a real-time manner to the SmartDP.DMK platform. On the platform, client data is organized and structured.

Data will be managed and processed on the DMP platform and analysis results will be output in a real-time manner so as to facilitate the formulation and adjustment of strategies during the operation stage.

Data Analytics and Strategy Formulation

Visiting data is analyzed through the data uploaded in a real-time manner. After TalkingData's data labels and client data are connected, client bases may be analyzed in a more detailed way.

The device information of the visiting client bases and competing projects (e.g., the house for sale across the street) was acquired and compared to the data of TalkingData. Population differences were compared in terms of basic population labels, device labels, online behavior preferences, offline traces, and distribution of work, residence, and amusement to guide the adjustment of competitive strategies. Also, a specific competing project was subject to in-depth analysis according to whether the competition was strong or weak. First, we explored the differences of a competing project and one of TalkingData's client projects in terms of secondary offline preference label dimensions (e.g., a person who likes to play tennis). Also, the offline location of a client for a single competitive product was targeted to assist client interception offline and avoid random customer visits.

Historical population migration analysis is done to observe the changing characteristics of regional populations and to predict future trends in combination with the current client analysis. Also, by combining the characteristics of regional population development, we may conduct a benchmarking analysis of unconnected areas so as to select areas with more competitive development strength. By gaining insight into population attributes within the region, we may formulate a targeted marketing strategy.

In terms of marketing, suggestions on marketing adjustments are given according to the big data. In terms of offline marketing, TalkingData combines self-owned offline location data and analyzes the regions where offline client bases occur most. It then concentrates offline client base activity, demonstrates the effect of offline promotion, and adjusts the main direction for future offline promotions based on the above.

Advertisement release assessment

During the process of marketing, both online and offline advertisement release strategies in the early stage were assessed. Through an analysis of app behavior preference of visiting clients, the installation rate of the actual visiting clients in the media selected was calculated and the effectiveness of online media selection was assessed. For offline billboard placement in traditional marketing channels, the coverage and capture rates for offline advertisement release were measured through matching of potential clients and visiting populations covered by GPS data. Both marketing and release strategies were promptly modified and the effect was continuously optimized according to whether the advertisement release assessment result was good or bad.

Development of lookalike populations

TalkingData introduced a data science team into the project and regarded marketing targets as seeds. After data (e.g., MAC address) was connected, the Loookalike algorithm was used to develop similar populations. Then, further marketing efforts were formulated to acquire clients through data analytics and business insights.

Here are the application contexts:

Diagnosis of operation health of shopping mall

Based on TalkingData's 3A3R indicator system for commercial operation (Awareness, Acquisition, Activation; Retain, Revenue, Refer), detailed indicators including client acquisition, retention, and activity were analyzed; problem indicators and their interaction were separated level by level and operating problems were diagnosed. For example, in the daily indicator monitoring, we found a warning of decline in monthly customer flow, marketing rate, and monthly sales volume, which is shown in Figure 7-4.

Indicators may be further mined and analyzed on platforms. It was discovered that there was an obvious fall in client traffic on business days. It was the same case with percentage of new clients, percentage of active clients, client traffic conversion rate, and average customer expenditure. This indicates that obvious risks occurred to the operation of the shopping mall in recent days. It was discovered that the lost client bases were mainly young and middle-aged female workers who were wealthy. And

based on the data analytics of TalkingData, these client bases obtained an obvious improvement in interest labels such as food, personal beauty, and life. However, the stores and campaigns of the shopping mall failed to satisfy such demands, thus causing the overflow of consumption. For these client bases, the shopping mall formulated a targeted promotion plan and combined app marketing and offline campaign in cultural, food, life, and beauty products. For example, coupons were offered for certain food and beauty products through in-app ads or via WeChat alerts. During the period of the campaign (within 28 days), 15.3% of lost clients were regained and 29.2% of inactive clients were activated. The client flow increased by 7.8% and the sales volume grew by 5.1%.

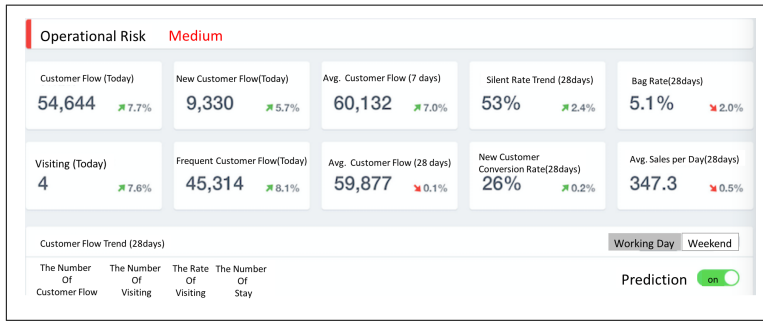

Figure 7-4. Customer flow analytics (figure courtesy of Yifei Lin)

Analysis of competitive product overflow and business district

Normally, a commercial real estate (i.e., shopping mall) prioritizes the composition, assets, and consumption capability of the populations living within three or five kilometers and then highlights visitor flow and overflow of surrounding shopping malls that sell competitive products and the characteristics of their visitors. In order to provide data support for business pattern layout, brand positioning, and customer management, the visiting client bases of competitive products were targeted using TalkingData's offline population distribution data and City Map; client flow and rate of overlap with competitive products were calculated and the flow and external consumption characteristics of client bases were analyzed. Also, competitive client bases and the potential client bases within the business district were compared through portraits for all populations and for competitive products within the business district.

Figure 7-5 shows that the shopping mall ("Project") has a high rate of overlap with competitive malls B, C, and D and shopping malls E and F are competitive. Through an analysis of client base characteristics and business district client base for the most competitive mall, B, it may be found that the Project and mall B are seizing young, medium-end client bases, and only capture 17% of the middle-aged high-end client bases in the business district. Thus, there remains a large space for attracting such client base.

Through further analysis of such a client base and in combination with data from TalkingData, it was found that household consumption dominated in these client bases, followed by female and business consumption. Thus, a series of adjustments were made including targeting release, business pattern optimization, and brand-level adjustment. As a result, high-end potential client bases began visiting the shopping mall once every week, up from the previous once every two weeks, and the ratio of such client bases increased from 10% to 13%.

TalkingData has gone deep into real estate enterprises, formulated the goal of business application and value creation, and rapidly formed a complete ecology of big data solutions. It has created practical applications and value for clients. With gradual accumulation of data and gradual perfection of the basic platform, big data will make the traditional real estate industry more vigorous and promising, and the further combination of data and contexts will create even broader value for real estate developers.

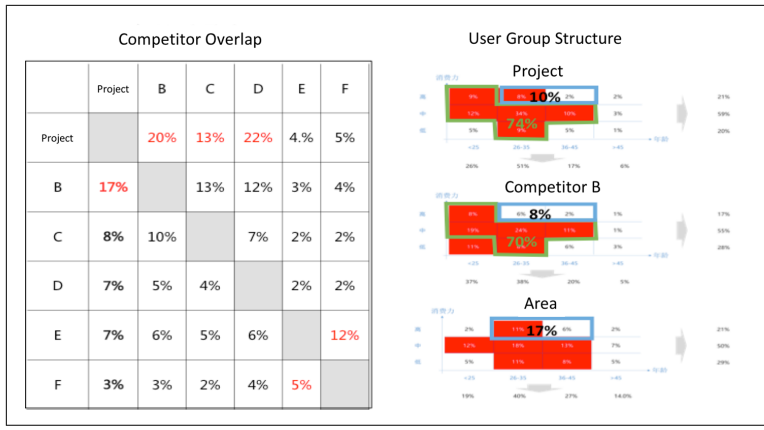

Figure 7-5. Customer structure analysis of Project and competitors (figure courtesy of Wenfeng Xiao)

Conclusion

The era of smart data has arrived, whether you have realized it or not. SmartDP's advanced technical platform can help enterprises respond to the challenges smart data presents in terms of data management, data engineering, and data science, while building an end-to-end closed loop of data. As we've discussed, SmartDP can provide flexible and scalable support for contextual data applications with agile data insight and data mining capability. TalkingData users have demonstrated that SmartDP can also greatly reduce the obstacles they encountered when transforming to a data-driven model, obstacles related to personnel, workflows, and tools for data acquisition, organization, analytics, and action. SmartDP ultimately improved their ability to drive contextual applications using data and explore commercial value, thus making them smart enterprises.

About the Authors

Yifei Lin is the cofounder and executive vice president of TalkingData, in charge of Big Data Collaboration with Industrial Customers. In this role, he focuses on Big Data Collaboration with enterprises from the finance, securities, insurance, telecom, retail, aviation, and automobile industries, helping traditional enterprises discover business value in mobile big data.

He has over 15 years of development, counseling, and sales experience, as well as 12 years of team management experience. He served as the General Manager of Enterprise Structure Counseling and General Manager of Middleware Technical Counseling for the Greater China Region for Oracle, the Senior Manager of the communications industry technical division for BEA, and the Senior Structure Consultant for Asia Info. He has also worked with several major Chinese banks (CCB, CUP, ICBC, SPDB, etc.), the three major telecom operators, large-scale diversified enterprises (China Resource, Haier), major automobile companies (FAW, SAIC-GM), and major high-tech enterprises (Huawei, ZTE).

Xiao Wenfeng is the CTO of TalkingData. He acquired a master's degree from Tsinghua University, and has worked in software development and development management for major companies such as Lucent, BEA/Oracle, and Microsoft. He joined BEA's telecom technical division in 2006, worked on the development of WLSS 4.0 (SIP signalization container based on WebLogic) as an architect and a core developer, and also led the development of BEA's first ISM full-service client project.

In 2008, he joined Microsoft to lead quality assurance for BizTalk middleware servers. In 2013, he joined Qihoo 360 as the lead of the PC Cleaner/Accelerator product division, and has managed the production, technique, and operations for multiple product families including the 360 Cleaner, and has applied for over 11 technical patents. In 2014, he joined TalkingData as the CTO and leads the development of all production lines.