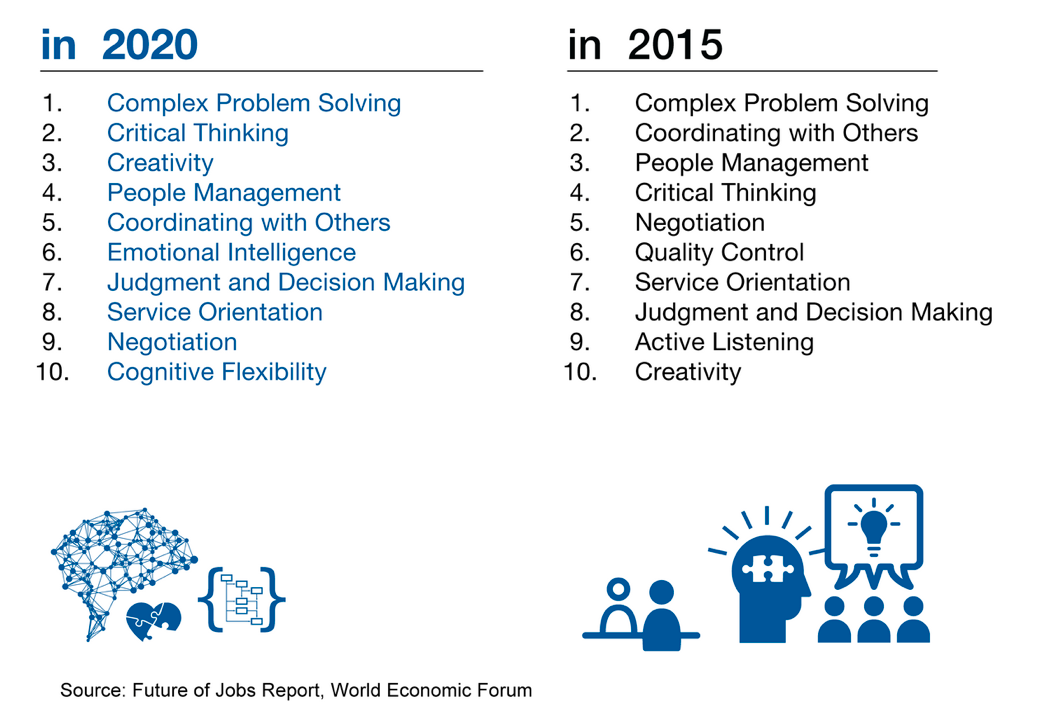
1. **Project Overview**
2. **The incoming Challenge of Origination**
3. **Strategy**

***Project Development***

According to the world trend, World Economic Forum indicates that we need to develop new strategies in response to The forth industrial revolution. Personnel training is definitely one of the major tactics. Please see Figure X for more detail.。From Figure X we should know, **Complex Problem Solving** is the most important personnel characteristic in 2020，其次為**Critical Thinking**, **Creativity**。Data Scientists-the sexiest job of 21 centuries, which characteristic also focus on Complex Problem Solving, Critical Thinking and Creativity. “**Data Scientists Training Project**” is initiated to evolve these trends to support CTBC to become Taiwan Champion, Asia Leader and the Best Financial Institute for Chinese around the world.



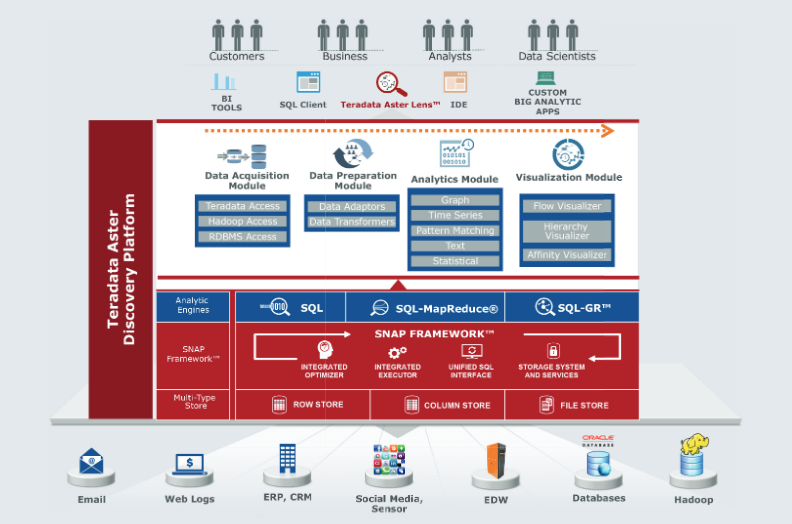
In order to achieve this vision, we start this investment to build a Big Bata Discovery Platform and cultivate many internal data analysts and data scientists in following dimensions:

## Build Internal Big Data Discovery Platform

For building the Big Data Discovery Platform, we developed three processes to implement and verify the performance of Big Data Platform. Finally, we chose **Teradata Aster Discovery Platform** to play the key role in Big Data Analysis in CTBC. Before talking about the process for building internal big data platform, here are the Teradata Aster brief introduction and architecture below.

**Teradata Aster is an analytic platform that embeds MapReduce analytic processing with data stores. The main advantages are:**

* Embedded MapReduce analytic processing and unique SQL-MapReduce® framework
* Massively parallel data store for multistructured data
* Intuitive tools and SQL-MapReduce libraries for rapid analytic development



The picture above is Aster Big Data Discovery Platform Architecture, from this architecture, we could understand there are various kinds of structured data and unstructured data, including Email, Web Logs, ERP, CRM, Social Media, EDW, Databases and Hadoop platform. Aster Database could process each kind of data in different ways like Row Store, Column Store and File Store. Beside, we could find it is a hybrid engine which include SQL Engine, SQL-MapReduce Engine and SQL-GR Engine. SQL-MapReduce is a big feature for Aster, it combines MapReduce advantage which could parallel process big data, it also packages MapReduce algorithm to SQL-like forms which could let SQL users easily analyze data in the familiar way and understand the efficiency of big data tool. Every Packaged SQL-Map Reduce Function are already developed by Teradata engineer, the Aster users just have to use SQL Language to call the Function name and do not need to know how to write Map Reduce program. For the analytics layer, Aster provides Data Acquisition Module, Data Preparation Module, Analytics Module and Visualization Module, the four modules could match the main big data analysis cycle, such as data preparation, data preprocessing, data analysis and visualization to discover business insight. For Analytics Module, Aster has 100+ pre-built SQL-Map Reduce Function, including Path Analysis, Text Analysis, Graph Analysis, Statistics Analysis, Clustering Analysis, Association Analysis, Data Transformation……, each main analytics function module has various sub function could be flexibly used together.

After the introduction of Teradata Aster Discovery Platform, now we further see the two processes of building internal big data platform as follows:

* **PoC**

We invited Teradata Big Data Analysis Team members to implement the PoC through data analytic cycle on **Teradata Aster Discovery Platform**, including requirement collection, data preparation, data cleansing, data analysis, visualization and final presentation. Although we focused on rick management at the beginning, we founded using same analysis functions but used in marketing view will show different results. It was an interesting idea, letting us come out many analytic topics and benefit to different departments including **CRM, Credit Card and Risk**.

* **Projects**

After finishing the PoC, we started to integrate Big Data Platform with current operation databases, including SAS Enterprise Guide and Teradata Warehouse and come out a new Project “**Data Scientists Training Project**”. Unlike general Big Data Analysis Project, we focused on letting many colleagues to be familiar with the new Big Data Platform and know how to use different algorithms to solve current business problems. For us, we emphasize to popularize the internal Big Data culture rather than only completing analysis topics which probably were done by vendors. The project execution process as follow:

* **Hardware and Software Installation and Setting**

Teradata project members helped to prepare and establish the Big Data Environment, including install software on four physical machine, install analytic function, and design encryption program for different data source connection to protect confidential information.

* **Big Data Platform Standard Training**

Teradata project members helped to design the Teradata Aster Discovery Platform training plan for CTBC colleagues who were assigned to be data analyst/data scientists.

* **Big Data Scenario Implement Training**

Teradata project members helped to lead CTBC colleagues to learn how to implement the following topics, during this training period, we can learn the spirit of many kinds of Big Data algorithms and understand the tips to discovery insights, the details will be discussed in **Attend Big Data Analysis Training Courses** paragraph:

1. Infection Path Analysis on Risk Customers
2. Connection Graph Analysis on Merchants
3. Relation Analysis between Balance Fluctuation and Default
4. Text Analysis on Customer Complaints for Lexicons

* **CTBC Internal Hackathon competition**

After CTBC colleagues learned all the Big Data Analysis Training Courses, for verifying their accomplishment, we host an internal hackathon competition. Everyone who joined this project and learned all the training courses should attend this race, attendee should be divided into six groups and took a limited time to use Teradata Aster discovery platform to analyze the topics which they concerned in their department. Further detail will be discussed in **Project Achievement** paragraph.

## Build Data Analysis Teams from Risk and Marketing sides

* **Six groups from different departments**

To cultivate internal data analyst and data scientist, we divided about thirty people into six groups, two groups from risk department, two groups from credit card department and two groups from customer relationship management department(CRM). There are two roles in each group, one is project manager, the others are data analysts. Everyone should use personal talent to cooperate with each other.

* **Benefits**

The benefits for three departments involved in are using big data analysis techniques to create three business values in different views: from Risk side, we could know the business focuses on Risk management, the objects are discovering more potential risk customers and knowing how to efficiently manage risk occurrences; from credit card side, we could know the business focuses on marketing of many kinds of credit cards, the objects are discovering more patterns of customer behavior to capture customer interests; from CRM side, we could know the business focuses creating excellent customer experience and discovering more potential VIP customers.

## Attend Big Data Analysis Training Courses

Teradata project team designed two levels of training courses, **basic** courses designed to guide CTBC colleagues to be familiar with Teradata Aster Big Data Platform. Here is the brief introduction for each courses.

**Basic**

1. **Introduction to Teradata Aster**

This is the first class of series big data courses. The course is designed for Database Administrator, System Analyst (Data Owner), System Administrator and System developers. From this course, we could know the basic introduction of Teradata Aster big data platform, besides, we could quickly understand some Aster use case from domestic and foreign experiences.

1. **Analytics and Data Mining Introduction**

This course is designed for Database Administrator, System Analyst (Data Owner), System Administrator and System developers. From this course, we could start to understand various Aster SQL/MapReduce Function. The course focus on four Analytics Function Modules, including Graph Analysis, Text Analysis, Association Analysis and Path Analysis. Apart from introduction of each Analytics Function Modules, we also could learn the real case application from these four function modules.

1. **Data Analysis & Aster System Management**

This course is designed for Database Administrator, System Analyst (Data Owner) and System Administrator. From this course, we could know the management skill of Aster, such as authentication design, database resource management.

1. **Teradata Aster AppCenter Introduction & Development**

This course is designed for Database Administrator, System Analyst (Data Owner) and System Administrator. From this course, we could understand the function of Aster App Center- a visualization tool which constructed on Aster platform. It could show many icons like app, every icon packages a set of Aster SQL Scripts which are designed from Data Analyst or Data Engineer. For managers or business users who need not to develop analysis program, they only have to click the packaged icons and set the schedule, the analytics report will automatically show up on the screen.

1. **Tableau Training**

This course is designed for System Analyst (Data Owner) and System Administrator. From this course, we could understand the function of Tableau- a famous visualization BI tool. Apart from the basic introduction of Tableau, we could further understand many scenarios by Tableau.

**Advanced**

There are four advanced courses that are suitable for bank industry in Taiwan, the goal is to driving CTBC colleagues to learn the core analytic ideas and related algorithms. From completing these four implement topics, we could image everyone who join this project can have more creativity to figure out more big data analytic topic for themselves. Here is the brief introduction for each courses.

1. **Infection Path Analysis on Risk Customers**

This course is designed for each data analysis team member to deeply know how to implement Graph Analysis. “Graph Theory” is a methodology which could be used to identify social network. We use “Graph Analysis” methodology to build Cash Flow Communities by customer deposit data and figure out the observation period and the performance period, try to find someone who has cash flow relationship with high risk customer during observation period. For verifying the result, we further discover the target customers who are affected by the high risk customer and also transform to high risk during performance period. At last, we summarize the reason and pattern why deposit relationship may influence someone in the same community.

These are the related Graph Analysis Functions below:

* Pagerank
* Betweenness
* Closeness
* Local Clustering Coefficient
* Modularity
* nTree

1. **Connection Analysis on Merchants**

This course is designed for each data analysis team member to deeply know how to implement Association Analysis. “Market Basket Analysis” is a method that figure out which items are often bought in the same order, and try to discover the best product sets to drive marketing value. We applicate “market basket analysis” in different way, analyze credit card statement data to discover which merchants are visited in the same credit card holder, try to figure out the merchant connection. At last, we summarize the potential merchant relationship sets which could be helped to come out marketing strategies.

1. **Relation Analysis between Balance Fluctuation and Default**

This course is designed for each data analysis team member to deeply know how to implement Path Analysis. “Path Analysis” is the method which often used in website analysis, try to figure out customer frequent path on website. We applicate Path Analysis in a new way. Our analysis target is the changed figure like deposit balance instead of a physical object like website’s clicked items, we try to figure out the balance fluctuation by path analysis method. At last, we summarize the frequent balance fluctuation path, also analyze the relation between the path and default. After finishing the analysis topic, we could come out strategies in advance before the customers who enter the same pattern in the future.

1. **Text Analysis on Customer Complaints for Lexicons**

This course is designed for each data analysis team member to deeply know how to implement Text Analysis. “Text Analysis” or “Text Mining” is a famous method to analyze unstructured text information like blogs, forum from social media. The goal is to transform the unstructured data to structured data that can be easily understood by text segmentation, key word extraction, sentiment analysis, text classification. The Text Analysis training course is divided into two parts, one is **new word discovery**, the other is **text classification by supervised machine learning method.**

* **New Word Discovery**

New Word Discovery is an important concept in text mining. It is not easy for Chinese word identification by segment method because of new words invention incessantly like brand, 3c product. Also, Chinese sentences could not be recognized by blank space. For resolving these problems, we use nGram method to find the sequential frequent word sets, if the word has never shown in the dictionary we built before, it can be directly put in the dictionary.

* **Text Classification by Supervised Machine Learning Method**

Machine Learning is a popular skill in data mining, it is divided into supervised and unsupervised method. Supervised machine learning means we have history data set that has already tagged classification results, system can learn the history data and predict the future data which do not have answer. Unsupervised machine learning means we do not have any tagged answers and directly do classification by math algorithm. For Supervised Machine Learning method in text mining, we use Naïve Bayes to predict the classification result on social media articles.

These are the related Text Analysis Functions below:

* TextTokenizer
* TF-IDF
* Ngram
* Sentiment Extractor
* Naïve Bayes

## (4) Internal Hackathon Competition

After finishing two levels of Aster Big Data Training courses, each team must use the Algorithms they learned before to finish the Big Data Analysis topics by themselves in limited time. Also, they should attend the final presentation and perform the Big Data Analysis result. Further detail will be discussed in nextparagraph.

***Project Achievement***

## Internal Hackathon Competition

There are six big data topics and key concepts for internal hackathon competition, each team members did their best to finish the race. Here is the detail below:

1. **The Butterfly Effect of Risk** (Department: Risk Management)

* **Analytics background and Target:**

CDIA is a contract that help customer to solve the debt issue in a different way. The figure of applicants has grown annually; we also find someone who break their promise. For resolving the growth of break rate, we try to use big data analysis method to discovery the break path of customers and expect to find someone who has cash flow relationship with them also has the probability of promise breaking. In Addition, we further observe the payment channel path, trying to decrease the human resource effort.

* **Data Using:**

Customer bucket records in every month, Customer CDIA History Records, Customer Deposit History Data.

* **Aster Algorithm:**

Path Analysis Module, Graph Analysis Module

* **Analytics Result:**

We discover some frequent break path and find pattern from someone who has been infected by the promise breaking customers. In the future, we could predict and decrease the break rate of CDIA contract.

1. **The Consumer Decision Analysis of Car Buyer** (Department: Credit Card Marketing)

* **Analytics background and Target:**

The reason of focusing on the behavior of car buyer is- car is a high-priced item and after buying the car, there are many related service they will need in the future, such as card debt, car insurance……, our target is creating the Consumer Decision Modelto analyze the behavior of consumer in different views that could help to recommend the right product on the right time.

* **Data Using:**

**Internal Data-** Customer Card Statement, Merchant Information, The Event That Car Buyers focus on, Car Buyer’s Profile.

**External Data-** Taiwan PPT Online Billboard, Taiwan Mobil01 Forum.

* **Aster Algorithm:**

Path Analysis Module, Graph Analysis Module, Text Analysis Module, Association Analysis Module.

* **Analytics Result:**

We discovered some frequent behavior path of car buyers that would help to understand customer requirement in advance. Besides, we found the proportions of insurance and car fixing are the highest, we could increase the opportunities with merchants or travel agencies.

1. **Customer Churn Path Analysis** (Department: Credit Card Marketing)

* **Analytics background and Target:**

In the past, we were passive to understand the churn reason of credit card customers. By using big data analysis tool, we expect to find the frequent churn path in advance, also try to figure out the right ways to contact customers.

* **Data Using:**

**Internal Data-** Customer Card Statement, Merchant Information, The Events That Credit Card focus on (customer adhesion degree, costs incurrence, marketing contact……), Credit Card Holder Profile.

* **Aster Algorithm:**

Path Analysis Module.

* **Analytics Result:**

We discovered about four thousand churn paths and summarize some frequent path. Besides, we further figure out some special paths which are representative, all the path results could be referenced by decision makers.

1. **Customer Life Circle Discovery Analysis** (Department: CRM Marketing)

* **Analytics background and Target:**

In the past, we found many customers’ life cycle are far away from the original bank in our bank. It resulted in the difficulties of marketing management. For solving this issue, we try to use big data analysis methods to build customers’ real life circle and the working circle to discover the right contact point. It could be a strategy to enhance working efficiency for customer relationship manager.

* **Data Using:**

**Internal Data-** Customer Card Statement, Merchant Information, ATM Transaction Detail, Credit Card Holder Profile.

* **Aster Algorithm:**

Graph Analysis Module.

* **Analytics Result:**

We successfully built many customer groups and further filter someone who is not close to our bank. Through deeply understanding customers’ life circle, we could be more close to our customers.

1. **Graph Analysis of Group Leader** (Department: CRM Marketing)

* **Analytics background and Target:**

The thought of topic is from “Group Buy”, it means someone who actively provides the information of buying the product and lets more people buy it together for the special prize. In General, this role has highly influence on his surroundings, he could have strong social network. We use this concept to implement Graph Analysis on the customers who have payroll transfer account and expect to let group leaders’ social influence enhance the distance of customers and our bank.

* **Data Using:**

**Internal Data-**CustomerDeposit Transaction Data, The Profile of Customers who has Payroll Transfer Account.

* **Aster Algorithm:**

Graph Analysis Module.

* **Analytics Result:**

We successfully find the group leaders and deeply increase the contact point through special marketing activities. Besides, we could start from the key group leader to attract other customers by MGM(Member Get Member) strategy.

1. **Potential Risk Customer Behavior Analysis** (Department: Risk Management)

* **Analytics background and Target:**

In the past, we used traditional variables that are not focus on different views of customers, it may cause the weak prediction of traditional model and is not easy for marketing application. We expect to use big data tools to analyze three kinds of customers, like **marginal high risk customers**, **thin file customer of deposit account** and **loan customers**. We try to discover the behaviors of these three kinds of customers, and could prevent risk in advance.

* **Data Using:**

**Internal Data-** Customer Card Statement, Customer Deposit Transaction.

**External Data-** Taiwan Mobil01 Forum.

* **Aster Algorithm:**

Path Analysis Module, Graph Analysis Module, Text Analysis Module, Association Analysis Module.

* **Analytics Result:**

For **marginal high risk customers**, we focus on the credit card holders who have broken the promise, finding some new management strategies to this kind of customers, we even could come out some new rules to decrease the target customer of risk management. For **thin file customers**, we discover their relationships, using different views to identify potential risk customers. For **Loan Customers**, we find some unusual behavior patterns that could be alerts to prevent more risk in the future.

1. **Significant Benefits**

***Significant Benefits***

## Financial Dimension

**此章節與貴單位內部專案後的評量方式有關**

## Qualitative Dimension

With this “**Data Scientists Training Project”**, the benefits could be found in different dimensions, it’s an overall enhancement for the Big Data culture in CTBC.

1. Big Data Environment Establishment

In this Project, we built Teradata Aster Big Data Discovery Platform in CTBC, with its’ parallel architecture and rich data mining modules, we can image there will be more Big Data Analytics Topics in CTBC.

1. Data Scientists cultivation

In this Project, we design three levels of training, including basic courses, advanced courses and hackathon competition. Although everyone was not familiar with big data analytics methods and didn't know how to use the Aster Big Data Discovery Platform in the beginning, during four months training and real experience, six group member already have overcome the difficulties even can cooperate with each other to finish the outstanding performance in the end in final hackathon presentation. From these experience, we successfully cultivate many potential data scientists, it could be a big fulfillment in CTBC.

1. Six hackathon big data analytics topics

From this hackathon experience, we have finished six Big Data Analytics Prototypes which will be easily to be implemented in the production environment and automatically operated in the future.

1. The Cooperation of Risk Management, CRM Marketing and Credit Card Marketing Departments

For Big Data advanced application, we finally should analyze complete life cycle of all customers to create customer single view. For the first cooperation with each different department, we start to know the importance of data integration, and will expect to using more kinds of data to create business values.

1. **Risk Control**
2. **Future Plan**

**可以參考：CTBC個金風險管理處-大數據分析實務啟動計畫**

1. **Why Win?**