Name: \_\_\_\_Chua Kian Ann\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Organization\_\_NCS\_\_

NUS MATRIC NUMBER :\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ *(if applicable)*

Kubera Financial Services Case Study

# Introduction

***Kubera Financial Services (KFS) Inc*** provides financial products in the fields of deposit taking, loans, investment services, insurance services and market intermediation for the offered financial products. ***KFS*** is an international organization, employing over 3000 financial service professionals and operating in over 80 countries. ***KFS*** serves multiple industries, including: telecom, commercial/business-to-business, construction, property management, commercial leasing, auto finance, media, healthcare, transport and logistics, manufacturing, technology and legal services.

The ***KFS*** current IT environment consists of a combination of client-server and mainframe platforms. ***KFS*** currently uses data warehouses and business intelligence tools for reporting on and analysing customer behaviour to better anticipate their needs.

Over the past few decades, the company’s profitability has been in decline. A committee comprised of senior managers was formed to investigate and make recommendations. The committee’s findings revealed that the main reason behind the company’s deteriorating financial position was their inability to increase product offerings. Therefore, ***KFS*** is planning to use data mined from customer interactions to categorize their customers into finely tuned segments. This will enable ***KFS*** to create increasingly relevant and sophisticated products to address the needs of each customer segment.

# Analytics Needs and Plans

In terms of analytics, ***KFS*** currently makes use of both ***descriptive and diagnostic analytics***. Descriptive analytics looks at past performance and understands that performance by mining historical data to look for the reasons behind past success or failure. Almost all management reporting such as sales, marketing, operations and finance, uses this type of post-mortem analysis. Descriptive models can be used, for example, to categorize customers by their product preferences and life stage. Diagnostic analytics are carried out as part of various business intelligence activities, such as performing queries to answer questions – for example, why last month’s sales target was not met. This includes performing drill-down operations to breakdown sales by type, industry sector and location so that the areas that have underperformed for specific types of products can be determined.

In the future, ***KFS*** is planning to make use of ***predictive analytics***. For example, predictive models exploit patterns found in historical and transactional data to identify risks and opportunities. Models capture relationships among many factors to allow assessment of risk or potential risk associated with a particular set of conditions, guiding decision making for candidate transactions. Also, by using prescriptive analytics it is anticipated that ***KFS***can further enhance the realization of its goals. For example, prescriptive analytics can be used in establishing price points for new products or services and sales lead assignment to ensure that a client is matched with the correct sales associate.

# Data Sources

The currently identified data sources include:

1. Traditional enterprise data from operational systems related to customer touch points such as:
   * ATMs.
   * Call centres.
   * Web-based and mobile sources.
   * Branches/Brokerage units.
   * Mortgage units.
   * Credit cards.
   * Debts including loans.
   * Volatility measures that impact the clients’ portfolios.
2. Financial business forecasts from various sources such as:

* News services.
* Industry data.
* Trading data.
* Regulatory data.
* Analyst reports (internal and competing companies).
* Alerts about events (news, blogs, Twitter and other messaging feeds).

1. Other sources of data such as:
   * Advertising response data.
   * Social media data.

The IT team members performed a categorization exercise of these data sources and prepared the following categorized list:

* **Structured data:** financial products data, customer portfolio, bank transaction records, currency data, credit claim data, debit details, loans and product quote data.
* **Unstructured data:** social media data, finance service application documents, call centre agent notes, social media feeds, twitter feeds, credit claim adjuster notes and incident reports.
* **Semi-structured data:** customer profile data, national finance policies reports, monitory authority policies data, foreign exchange rates data, webserver logs and e-mails.

# Identified Big Data Characteristics

The ***KFS*** IT team members have assessed the different datasets that are generated inside the company’s boundary as well as any other data generated outside the boundary that may be of interest to the company in the context of volume, velocity, variety, veracity and value characteristics.

## Volume

The team noted that within the company, a large amount of transactional data is generated as a result of processing financial services, selling new products and changes to existing products. Also, a quick assessment revealed the existence of large volumes of unstructured data, both inside and outside the company.

## Velocity

With regard to the in-flow of data, some of it was assessed as ***low velocity***, such as the claims submission data and the new sales of credit services. However, data such as webserver logs and credit product sales quotes were assessed as ***high velocity***. Looking outside the company, the IT team members anticipated that social media data and the foreign currency appraisal data would arrive at a fast pace.

## Variety

In pursuit of its goals, ***KFS*** will be required to incorporate a range of datasets that include customer records, products data, claim data, sales quote data, social media data, call centre agent notes, claim adjuster notes, currency trading reports, transaction data, webserver logs and e-mails.

## Veracity

A sample of data taken from the operational systems shows signs of ***high veracity***. The IT team attributed this to the data validation performed at several functional processing stages in the creation and processing of financial services. Looking outside the company’s boundary, a study of a few samples taken from the social media data and currency exchange data demonstrated increased levels of data validation and cleansing to make it ***high veracity*** data.

## Value

As far as the value characteristic is concerned, IT team members concurred that the company needs to draw maximum value from the available datasets by ensuring the datasets are stored in their original form and subjecting them to the most appropriate type of analytics.

# Current System Capabilities

Most of the ***KFS*** operational information systems utilize client-server and n-tier architectures. After surveying the inventory of IT systems, the IT team determined that none of the systems employ distributed data processing.

# Anticipated Processing Workloads

A Big Data processing workload is defined as the amount and nature of data that is processed within a specified amount of time. The processing workloads for the proposed ***KFS*** Big Data engineering initiative are of three major types (1) Batch Processing (2) Real-Time or Stream Processing and (3) Transactional or Data Store Processing. These are briefly described in the following subsections.

## Batch Processing Workloads

Batch processing is the general process of computing information from vast amounts of data by using a distributed cluster of machines, typically over the course of several minutes or hours.

## Real Time or Stream Processing Workloads

Streaming can process real-time data and events in milliseconds, allowing computations that would be too complicated with messaging queues alone. There are several solutions that support different levels of throughput and guarantees that the data is processed, even when machines or the network fail.

## Transactional or Data Store Processing Workloads

Once data has been processed by the streaming or batch computations, it needs to be stored in a way that can be quickly accessed by a data scientist. While file systems are designed to store data durably, databases organize data in a way that minimizes unnecessary disk seeks and network transfers to provide the quickest response to queries.

# Key Business Challenges

When defining Big Data engineering platform-based solutions, many existing business capabilities can be enhanced when more and varied data becomes part of the information architecture. ***KFS*** is considering the following:

1. **Enterprise Modeling and Analytics Platforms:** ***KFS*** wants to build data reservoirs that become storage for all operational as well as non-traditional data sources. Analytics experts will explore this data reservoir for many new financial use cases. They will explore data, and build and deploy analytics models that will reveal new business opportunities or improve the solution of existing business problems.
2. **Mobility and Location Based Services:** New finance offerings can be designed and offered to customers when the analytics engineer connects the buying behavior of a customer with location. For example, a time bound offer for a local restaurant that has a relationship with the bank can be made to a customer via their mobile device as they walk into a movie theatre.
3. **Customer Intimacy:** Every engagement with the customer can be a selling opportunity. Better understanding of the customer, their traits, how they like to communicate, services they consume, and their value to the business enables the right product to be positioned to the customer at the right time for the right price.
4. **IT Operational Efficiency**: A possible reason for embarking on extended architectures that include the Hadoop Platform is the need to move data staging and transformation to a schema-less platform for more efficient processing and leveraging of IT resources. IT operational efficiency is an initial justification that ***KFS*** gravitate towards when deploying Big Data engineering solutions.

**Question 1** *(refer also to Appendix A)*

The ***Kubera Financial Services (KFS)*** case study is described at appendix A. You are a Big Data engineer working for KFS. Answer the following questions from a ***Big Data Engineering Architecture and Design*** perspective:

1. Consider the key business challenges identified in section 7 of the KFS case study. Identify **three analytics projects that use big data processing** for data driven decision making and actionable insight engineering. Justify your answers by explaining how business benefits from those projects.
   1. Project 1 – Offer more competitive interest rates for Fixed Deposit
      * KFS can leverage on they are existing data such as customer portfolio, national finance policies reports, monitory authority policies data to have an analytics where there would be an attractive interest rate that KFS can offer to their customers
      * KFS can also use data from customer portfolios to analyze their risk appetite.
      * With more customers placing money into their Fixed Deposit accounts, KFS can use the money to invest in other investments to yield a better return.
   2. Project 2 – Ensure better User experience with the KFS system for customers.
      * KFS can leverage on data like social media feeds, twitter feeds and social media data on their competitors and advertising response data on how they promote on their social media and marketing.
      * KFS can also use data for the web-based and mobile sources to analyze the web applications logs to identify the type of mobile devices they are using. If the large majority of the customer is using a certain mobile device, KFS can partner with their to have a marketing campaign
      * As User experience can make or break the company, from on-boarding them to create an account, to using the system to buy finical products or perform transactions. Without this the customer retention will drop. The idea is to give the customer a pleasant experience, a sense of inclusivity.
   3. Project 3 – Optimal date time to release a new product.
      * KFS can leverage on data from the news services, alerts about events via social media or twitter feed, web application logs to understand the usage activity of their customer and to understand is there any breaking news that would shadow or undermine their new product launches.
      * From there they will be able to understand which set of customers are active at which period of time. For example, if KFS wants to launch a new saving account for children, they can use the model create to find the optimal window to launch it in order to have the maximum reach to their customers.
2. The KFS IT team follows a step-by-step approach to implement Big Data engineering solutions. List the key **stages or layers** that you would design to process the several data sources identified in the requirements of KFS?
   1. Data Source & Ingestion
      * The data will be obtained from KFS traditional enterprise data from their operational systems, financial business forecast and other source for example past advertising response date via survey and social media data.
      * As for the Ingestion they can consider Amazon Managed Streaming Kafka. The solutions are serverless which only requires minimum maintenance. So KFS engineers can focus on the analytics portion.
   2. Storage
      * The next step is to determine what type of storage to store the data.
      * As mentioned KFS wants to move data to a schema-less platform. As a start KFS can consider storage from major cloud providers from AWS S3 or DynamoDB for OLAP.
      * For the transactional data storage, KFS can also consider storage from AWS RedShift as their data warehouse for OLTP.
   3. Query
      * The next step is to determine what query language to use. KFS can consider using AWS Athena which uses SQL to query date out
      * This will reduce the learning curve of the the engineers in KFS and allows the engineers to get things done faster and KFS will be able to get their ROI faster.
   4. Compute
      * As the current infrastructure of KFS they do not have a distributed data processing
      * They can consider using Amazon EKS to host a Kubernetes Cluster and install Spark on it as their compute layer.
   5. Data Processing
3. In the big data engineering context, workload processing could be classified as either *batch*, *real-time* (streaming) or *transactional* (active transactional data store processing). In the KFS case study context, which of these workload processing design choices would be most suitable for each of the following scenarios? Justify your answers:
   1. Popularity Analytics: The popularity of a product is determined by finding out how many times the corresponding page of that product was viewed. The webserver creates an entry in a log file whenever a webpage is requested.
      1. This would be classified as a real-time workload because as the customer uses the website to view the product, you would like the latest data on what the customers are viewing and buying. With real time data streaming in, the management can take advantage of this data to either create a promotion for the items and push this items to other customers to increase sales.
   2. Sentiment Analytics: The IT team believes that the event stream processing model can be used to perform sentiment analysis on Twitter data to find the reasons behind any customer dis-satisfaction.
      1. This would be classified as a batch processing, The reason for this it because the event has already happened and having bad customer service/experience only happens to a small subset of customers.
      2. Through this batch processing, they can slowly analyze the data and fix the issue on their IT system
   3. Invoice Systems: One of their accounting systems processes all business invoices as transaction (example: billing or sales).
      1. This would be classified as a batch because invoice systems are usually billed at the end of each month, so real time streaming is not needs.
      2. As the invoices system are calculated so a batch processing would make sense.
   4. Fraud Analytics: The KFS data engineering wants to identify credit card fraudulent activities during payment process.
      1. This would be classified as a real time because to identity fraud, it would need to analyze each transaction on what the purpose of the transaction in real time to block any suspected transaction to protect their customers.
      2. It cannot be a batch processing because it will be too late to block and it will create a lot of overhead for any refunds to the customer.