Personalized Finance Planner

CSYE 7250 – BIG DATA ARCHITECTURE AND GOVERNANCE

FALL 2020

KAMBIZ HEYDARI

Madhumathi Prakash

NUID: **001197053**

Contents

1. Requirements

1.1. Functional Requirements

1. Business Requirements
    - When signing up for the financial tool for the first time the customer should be asked about their age, goals, salary, type of investment, family status and should be given the recommended option of the tool to do the recommendation and changes to the account.
    - The user should be getting quarterly reports sent to their dashboard as well as emailed to them.
    - As there is already an application in use for the assistance for the financial planning, an upgrade will be provided for the UI which will provide a variety of configuration options as it will be easier for them to not completely jump into the new tool.
    - Goal is to completely phase out the old tool and move the new tool.
    - There should still be the option to “check out investment professional” for older clients.
2. Business use
    - As someone in Personal Investing or as a analyst there should be a separate login for the case of pivoting and creating reports.
    - Will have analysis along with the results.
3. Data
    - Historical data is currently in the form of relational tables which requires migration to NoSQL as it is the easiest unstructured form of the data and accommodated big data methodologies.
    - Data Profiling would take care of the tables, their attributes and the data that they handle.
    - Since majority of the use for the algorithms and predictions for financing per individual needs and characteristics depends on historical data, there needs to be data transfer from the legacy new version
4. User Security

Authentication and Authorization Requirements play an important role while defining functional requirements for system. They provide user access information and provide security to the system.

- **Authentication**: It involves management of system access with respect to user. Authentication is a process by which you verify that someone is who they claim they are. This is extremely important when user signs back in because this is very sensitive information involving people’s personal finances. It is important to keep the data very secure.

- **Authorization**: Authorization is the process of establishing if the user (who is already authenticated), is permitted to have access to a resource. Authorization determines what a user is and is not allowed to do.

5. Blocking threats with Audit Tracking or Audit Logs

- Audit tracking is one of the functional requirements for a system which is significant especially with all the sensitive personal data, such as SSN.
- An audit trail or audit log catches an abnormal change or addition to information and it is "red-flagged," the better the response to mitigate against negative influences such as cyber-threats, security breaches, data corruption, or misuse of information.

6. Legal or Regulatory Requirements – It is to ensure that all necessary governance requirements can be met.

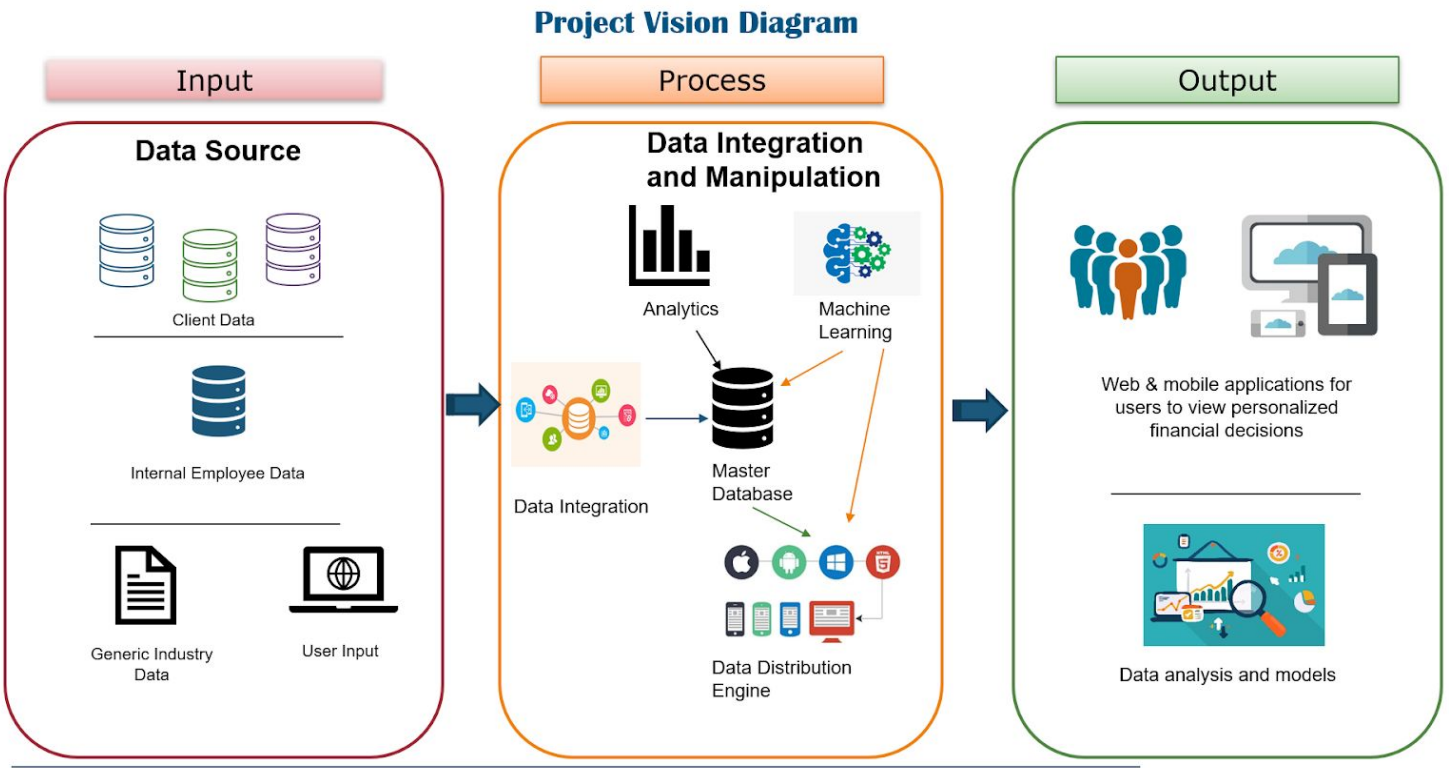
1.2. Non - Functional Requirements

1. Data requirements
   - No data loss in streams, data streaming, handle various data sources, historical data availability, updating the models, data sources and visualization layouts, scalability of the system, data privacy and security, and recovery in case of disaster.
2. First step is to break up the plan into smaller components aka a work breakdown structure (WBS). This way each task is broken down into a smaller portion.
   -
3. Reliability
   - System should be able to perform the functionality of predicting and investing based under any condition for a specific time.
   - Availability – the system should always be up and running as customers would access it at any time. However, there could be down time and for that reason there should be scheduled downtime and maintenance during the least affected time.
   - Failure rate – Since the system does with personal investing tailored to the specific user it should not be crashing but the down time can be after business hours if there is a need for correction since the need to invest in stocks or anything is not so dependent at that time. No single point of fail
4. Performance
   - As the user input is modified at the beginning while signing up or if user has a modification to their goal, the reaction of the response does not completely depend on that. The categorization and analysis with the algorithm should be instant as the investment path might need to change accordingly
   - The system can accomplish high availability through the use of data stores or marts.
   - Scalability in terms of architecture and support design.

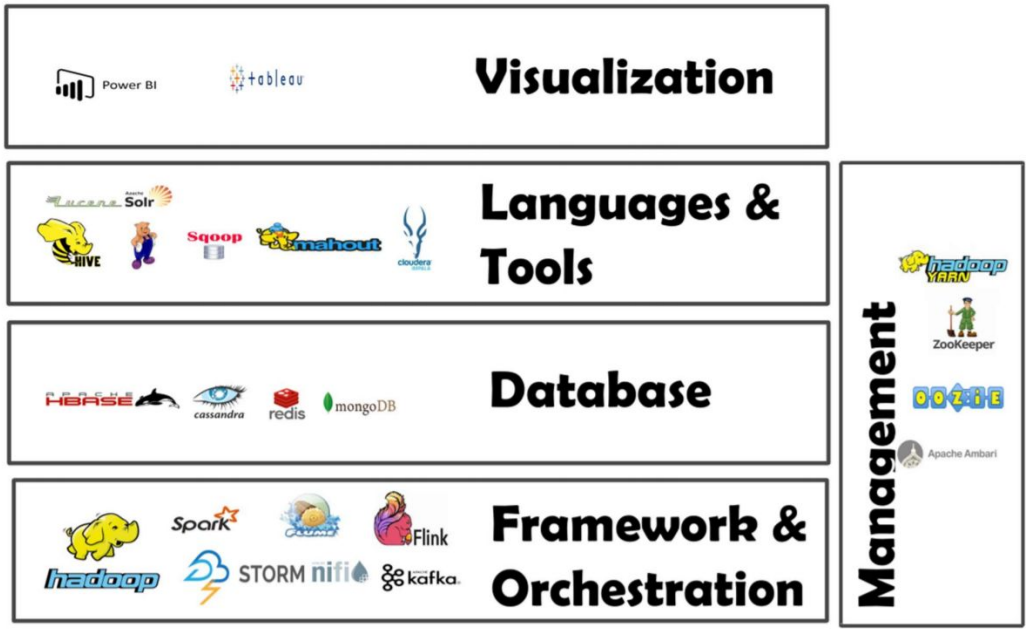
5. Security
    - All the data coming in must be backed up and be held in a secure system as it is personal client financial data.
    - No external personals should be getting access to this.
    - The data internally can only be seen by Personal Investing, analysts, Workplace Investing and each of these group of people can only access the set of data they need to see
    - Data must be encrypted.
6. Usability
    - Usability is key as there is already an application that is being used and if the placement and layout of the page completely changes, then it will be harder to retain those existing clients.
    - Must have well structured user manuals and informative error messages
    - Best way to test out usability of the application is to measure user satisfaction either through surveys or through eye tracking usability labs.

2. Overall Strategy and Architecture

2.1. Vision Diagram



2.2. Architecture



2.2.1 Visualization

Visualization is needed in the process step of the vison diagram and to analyze the user historical and current data by the analysts, Personal investor group and come up with modifications based on those. Also, to see trends and even display some statistics to the users in their dashboards so they can get an idea of what is occurring in the background and general statistics of clients with same criteria.

The visualization tool that would be best is Tableau. Tableau is a Business Intelligence tool for visually analyzing the data. The users, in this case research analysts, data analysts in the personal investing and work place investing areas, would create dashboards and reports. They would be able to share the dashboards and reports internally and only to people with access. The reports depict the trends, variations and density of the data in form of graphs and charts. Tableau can connect to files, relational and Big data sources to acquire and process data.

Some of the reasons to choose Tableau is that Fidelity Investments uses this visualization tool in high quantity across the organization, so it would be easier to find resources and not much training would be necessary. Also, the software allows data blending and real time collaboration which is key for the dynamic and high quantity of data. Tableau will prove to be ideal due to its ability to handle huge data sets that exists historically and incoming client data. Tableau will be useful for visually representing data which was analyzed and will give a clearer picture. In addition, a neat feature of the tool is the ability to connect to database and update based on real time data. There are two ways to connect Tableau to your database - Live and

Extract. If you set it as live connection, dashboard should get updated as your database is updated. Tableau is fast and has an easy to use intuitive interface.

2.2.2 Languages and Tools

These tools allow for distribution and consumption by systems and business applications.

The Cloudera tools from the US based company would be a perfect fit for this application. It allows for data engineering, data warehousing, machine learning and analytics that runs in the cloud or on premises It allows for developers, data scientists to build own algorithms.

It has Cloudera Data Science Workbench that allows for data scientists to experiment faster through the use R, Python, or Scala with on-demand compute and secure access to Apache Spark and Apache Impala. There is also the Cloudera SDX which lets fabric that makes multi-disciplinary analytics easier to develop and enables safe self-service access to all relevant data and increases compliance. It is essential to have a tool that not only allows for the use of real time data but also the ability to build own algorithm for personal financing.

Since Fidelity Investments has a high rate of data coming in as well as hosting data that already exists, it is important to have a tool that can withstand and work with all of that. In addition, the idea of this personalized financer is to make sure to look at historical statistics as well as new ones to determine based on the life goals and characteristics what would be the best fit. To do this, it is important to have a tool like Cloudera that allows not to only work in Premise but also if needed in the cloud and for the data scientists to manipulate the data easily.

Cloudera Data Warehouse delivers:

- Traditional data warehouse functions for BI reporting and modeling
- Data analytics
- On-demand self-service data access
- Interoperability with machine learning engines and algorithms for easier experimentation
- Hybrid choice to run on-premises, on public clouds or any combination
- Security, control and governance for diverse data and analytics

2.2.3 Database

For the personalized financial planner, it is critical to choose a database that can not only deal with the existing large amount of data, but also the continuous inflow of client data and some data from external source, which can be seen in the input part of the vision diagram. We are dealing data integration, evaluation, analysis, and interpretation of all source data. It intends to provide the ability to integrate, evaluate, and interpret information/data from available sources to create a finished intelligence product for presentation or dissemination to enable the ability to run machine learning processes on the integrated data and show analytics

or determine outcome. Also, to predict or separate the invested money by the clients appropriately based on all these characteristics.

Criteria/ Types of DBs	Key-value	Document	column	graph
Data volume	Large	Medium	Very large	large
volume of reads and writes	Large reads and writes	Medium	Very high	high
What is your tolerance for inconsistent data?	low	Low	high	High
Do you need high query capabilities to extract the data you need?	low	low	High	high
Does your application involve relationships/interactions between various entities?	yes	no	yes	yes

There is really a need for column family databases use case here since there is potential a large volume of data and the application will always be writing to the data base. In attrition there is geographic distribution over multiple data centers in North Carolina, Texas, and New Hampshire.

Data at hand is in terabytes and requires a highly scalable database. The database used is a combination of Apache Hive and HBase.

2.2.4 Framework and Orchestration

The framework that would be best for this project would be Hadoop. It is made up of file system and OS level abstractions, a MapReduce engine and Distributed File System (HDFS).

Hadoop's HBase database accomplishes horizontal scalability through database sharding. Hadoop is designed to be run on clusters of commodity hardware, with the ability consume data in any format, including aggregated data from multiple sources. This is important since the financial data coming in about a variety of clients can be coming from many forms of data.

Below are the advantages of Hadoop

1. Scalable

Hadoop is a highly scalable storage platform, because it can store and distribute very large data sets across hundreds of inexpensive servers. It also allows large companies like Fidelity Investments have applications on large amounts of nodes that are made by of high quantity of data, even in the exabyte level.

2. Cost effective

The main cost effectiveness about Hadoop is that it had the concept of scale-out architecture which allows to store all the data for later use if not needed immediately and this way saving costs. It could be costing thousands to tens of thousands of pounds per terabyte, Hadoop offers computing and storage capabilities for hundreds of pounds per terabyte.

3. Flexible

Hadoop enables businesses to easily access new data sources and tap into different types of data (both structured and unstructured) to generate value from that data. This means businesses can use Hadoop to derive valuable business insights from data sources such as social media, internal traditional user inputs, client data from the data forms in the application or clickstream data. In addition, Hadoop can be used for a wide variety of purposes, such as log processing, recommendation systems, data warehousing, market campaign analysis and fraud detection. In this case it fits our process of recommendation, prediction and actual decision making of how to split up the data according to user needs and modifying based on changes.

4. Fast

Hadoop's unique storage method is based on a distributed file system that basically 'maps' data wherever it is located on a cluster. The tools for data processing are often on the same servers where the data is located, resulting in much faster data processing

5. Resilient to failure

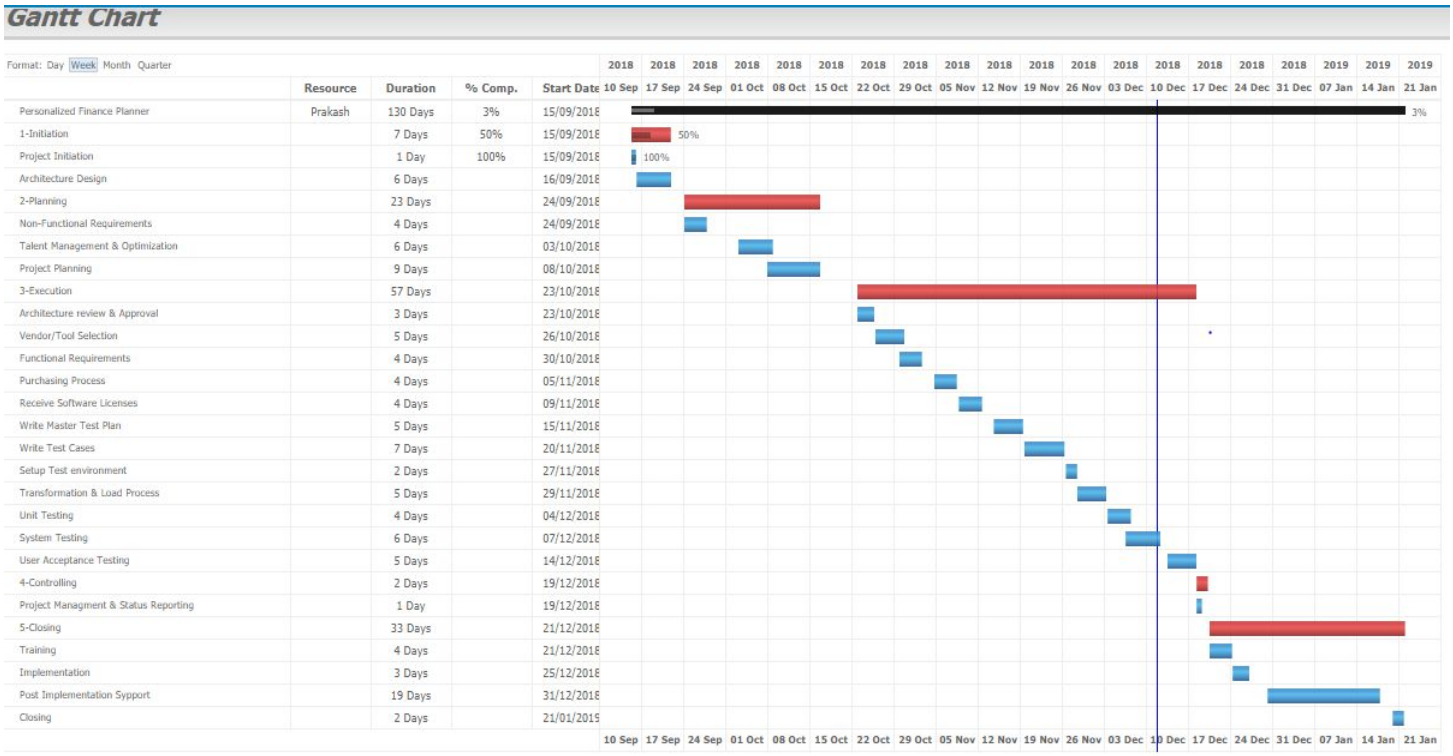
A key advantage of using Hadoop is its fault tolerance. When data is sent to an individual node, that data is also replicated to other nodes in the cluster, which means that in the event of failure, there is another copy available for use.

When it comes to handling large data sets in a safe and cost-effective manner, Hadoop has the advantage over relational database management systems, and its value for any size business will continue to increase as unstructured data continues to grow.

In terms of orchestration the use of Apache Flume is best as it can be integrated with Hadoop. IT is known for efficiently collecting, aggregating, and moving large amounts of log data (Hadoop Distributed File System (HDFS).). It has a simple and flexible architecture based on streaming data flows. It is robust and fault tolerant with tunable reliability mechanisms and many failover and recovery mechanisms. It uses a simple extensible data model that allows for online analytic application. Flume will transfer the data to HDFS.

3. Other Factors

3.1 Complete Project Plan



3.2 Issues and Risks

3.2.1 Issues

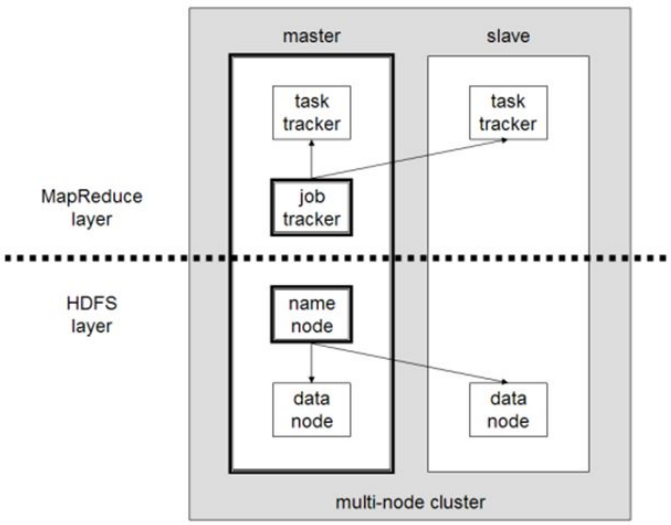
The biggest issues are:

- The quality of the data, the access and security of the data, and managing the current development and future cost of the application.

3.2.2 Risks

There should be talented data scientists, architects, support, and resources in general. Also there is the risk of tool malfunction or vendor backing out. Lastly there is also the data scalability risk from the terabyte amount of data.

3.3 Data collection



Hadoop has two main parts – a data processing framework and a distributed filesystem for data storage. The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. There is the Hadoop Distributed File System, HDFS, and it is used like a dumping place that holds all the data till it needs to be utilized and that is key with such a large amount of data of clients and all that information about certain age and life style group do not need to be used at all time for analysis. The key thing to remember is it stores data and you can pull data out of it, but there are no queries involved. It is MapReduce that processes the data with the help of Hive.

There will be the use of multiple machines, the work tends to be divided out: all of the data sits on one or more machines, and all of the data processing software is housed on another server.

The user inputs all the data in the system based on what their goals are, when they want to retire and current salary, and so forth. This data is stored in data blocks on the DataNodes. HDFS replicates those data blocks, usually 128MB in size, and distributes them so they are replicated within multiple nodes across the cluster.

After the collection of data Talend provides native and optimized code generation to load, transform, enrich, and cleanse data inside Hadoop without additional storage or computing expense.

When it comes to handling large data sets in a safe and cost-effective manner, Hadoop has the advantage over relational database management systems, and its value for any size business will continue to increase as unstructured data continues to grow.

3.4 Security

Security is essential to be managed between all the architectural levels.

- To stat out with, since there is Hadoop, which is a non-relational database, the security is a little tricky and needs to be kept up at all times.
- The high amount of data from external sources about other financial statistics can be malicious and additional security measures need to be taken. This can be down by validating if the source is trustworthy.
- Unethical IT specialists practicing information mining can gather personal data without asking users for permission or notifying them. So, it is important to give just the specific user access to certain data analysts in the personal investing section of the company.
- Access control encryption and connections security is essential.
- Have Information Barriers based on who requires the data? Some organizations don’t institute access controls to divide the level of confidentiality within the company.

- Recommended detailed audits are not routinely performed on Big Data due to the huge amount of information involved.
- Due to the size of Big Data, its origins are not consistently monitored and tracked.

There are three main categories of security issues: user authentication, data privacy, and data integrity**.**

1. User authentication

For the actual clients using the system it is important:

The use of simple passwords like abc is not make sure users select strong passwords and change them regularly. You can never be 100% sure people haven't tricked into revealing their password or if they are logging in to your company's systems from unsafe devices.

For the employees of Fidelity who are accessing the data:

Make sure you implement strong access and authentication tools and controls that will not permit users to access sensitive data unless the security of the device and network they're using can be verified. There can be the use of two factor authentication and also utilize the already existing practice of information barriers.

Each tool provides their own form of security like,

- Tableau: SAML and OAuth Authentication, Roles and permissions, Data and Network Security
- Hive/Hadoop: Kerberos authorization support, SASL Integration. Kerberos authorization support, Server-side/Client-side Configuration for Secure Operation with Thrift and REST gateway
- Talend: LDAP Authentication

2. Data Privacy

If you assume that determined hackers can breach your perimeter security, it becomes far more important to ensure they can't get away with anything of value if they do. This is crucial for data security. Make sure any sensitive data (whether stored or in transit) is hidden and is encrypted and only certain people can have decryption ability for the user historical and current data.

Moving from a traditional security architecture to a data-driven intelligence and response model is a major change with significant cultural implications. So make sure in terms of security everyone is trained.

Big data technologies weren't originally designed with security in mind. Using open-source technologies like Apache Spark and certain versions of Hadoop can help address this challenge.

3. **Data Integrity**

Build servers based on secure images for all systems in your organization's big data architecture.

As mentioned in the functional requirements it is important to ensure that auditing, maintaining, and analyzing logs are done consistently across the enterprise. To keep constant tabs on any potentially suspicious activity occurring on your network, you'll need an up-to-date security information and event management system such as Splunk or LogRhythm. It is used to identifying APTs. They collate information in real time from many sources and determine the threats the arise.

3.5 Scalability

The application should deal with any amount of scalability. Below is the issue and how to deal with them.

- There is an increase in number of data points or data storage

Hadoop: Hadoop allows for the distributed processing of large data sets across clusters of computers, thereby removing the data ceiling. It revolves around the idea of data scalability. IT deals with multiple hardware integrations as well.

Apache Flume: The performance moves up linearly by adding more resources to the system and it can increase the throughput. The way to do measurement in this tool is the number or size of events entering the system and having a load increase by adding resources as well.

Because of the definite increase in data over time it is great about Hadoop being no-relational and that it has the ability to store unwanted data separately. Hadoop allows big data analytics projects to power real-time customer interactions.

3.6 Management

The tool that can be used is Hadoop yarn. Apache Hadoop YARN sits between HDFS and the processing engines being used to run applications. It combines a central resource manager with containers, application coordinators and node-level agents that monitor processing operations in individual cluster nodes. YARN can dynamically allocate resources to applications as needed, a capability designed to improve resource utilization and application performance compared with MapReduce's more static allocation approach.

Apache Hadoop YARN decentralizes execution and monitoring of processing jobs by separating the various responsibilities.

There is the data availability, usability, integrity, and security of the data employed in an enterprise. A managed application and projects is made up of, a defined set of procedures, and a plan to execute those procedures.

People: An organized structure of highly skilled people have been considered for the data governance program.

The Stakeholders will also be an integral part of the process as they will be providing feedback and get regular updates on the progress of the program.

Process: The effective measurement of data governance program is a basic component of a successful program. Due to the sensitivity of the data, strict policies for the access of data has to be maintained. Regular backups, in case of data disaster can prove to be vital to success of the project.

There are a few items that are key in success and needs to be managed for this application:

1. Data will be combined from multiple data.
2. Implement Data Governance best practices.
3. Data Access: Increasing access to data while handling Data Security.
4. Analytic Prioritization: what set of data needs to be looked at first and what can be looked at later.
5. Ability to manage all the tools and software used

In terms of the management of the overall application:

- Load options and issues
- Generalizing keys for changing dimensions and aggregates/summaries
- Sorting data prior to the load
- Aggregation: building aggregates/summaries
- Exception processing
- Data Quality Assurance
- Data Publication
- Backup Strategies
- Capacity Planning
- New Administration Roles

Specific tool management:

Hadoop/Hive: There is the manageability of the data backup and recovery, performance tuning, data pruning, and data modeling

Apache Flume: Management and ability control data flows, monitor nodes, modify settings, and control outputs of a large system.

Overall it is important to have a centralized management point to monitor and change data flows, and the ability to dynamically handle different conditions or problems.

For management and monitoring, there should be:

- Legacy Data Extraction: Extracting data from Legacy Systems and other feeds
- Data Transformation: Transforming data into load record images
- Change Management: Managing data changes within the Legacy Systems
- Data Quality: Monitor and insure data quality