# PROPOSED SOLUTION & METHODOLOGY

## Product/Solution Overview

### Cloudera Data Platform

Cloudera Data Platform (CDP) is a data cloud built for the enterprise. With CDP, businesses can manage and secure end-to-end data cycles (collecting, enriching, analyzing, experimenting, and predicting with data) to drive data-driven insight and decision making. Valuable and transformative business use cases require a multi-stage analytic pipeline to process enterprise data sets. CDP empowers businesses to derive value from large, complex, distributed, and rapidly changing data; and compete in the digital transformation era.

***An Integrated Data Platform***

CDP provides an integrated data platform that creates agility across business lines and facilitates efficiency and security in IT, enabling the entire organization to be more productive. With organizations reacting quickly to changing business needs, CDP provides essential advantages:

* Designed for data engineers, data scientists, BI analysts, developers, and enterprise IT
* Cloud-native service that is easy to use and designed to be secure automatically
* Best in class analytics and integrated data cycle
* Self-serve and custom analytics
* Public cloud and on-premise

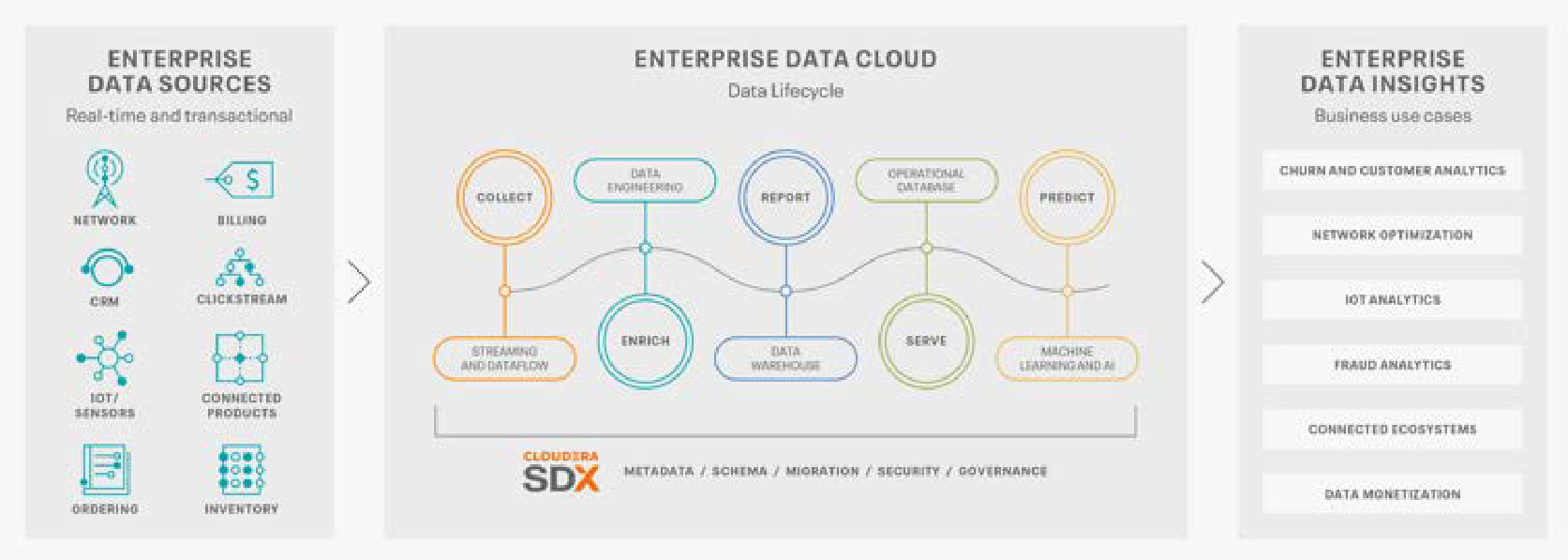
CDP allows enterprise IT to embrace these seemingly opposing forces as it delivers enterprise data cloud capabilities.**

Figure 1 - Cloudera Data Platform Overview

***Any Cloud: Public, Private, and Hybrid***

CDP is optimized for hybrid clouds and provides the same data management and analytics capabilities on private and public clouds. Enterprises can provide data management and analytics solutions in the environment of their choice. CDP separates data management from strategic infrastructure, so companies can move data and applications from one domain to another without rewriting applications and training personnel. These environments have varying degrees of scalability, flexibility, and cost-efficiency. CDP ensures that the business can choose the most appropriate setting based on use cases (data privacy regulations, type of analytics required, etc.), business objectives (speed of deployment, cost efficiency, etc.), and technical considerations (source and location of data, required elasticity)., etc.).

***Data Lifecycle Integration: From the Edge to AI***

Integration between streaming, analytics, and machine learning (data cycles) is a prerequisite for almost all use cases in data-driven businesses. With CDP, enterprises can connect disparate workloads to develop applications on a single data management platform to meet the needs of current and future workloads. By enabling enterprises to control data across environments, enterprises can master the data cycle to gain insights that can increase productivity and continue their transformation into a data-driven organization.

***Secure and Governed***

Cloudera's Shared Data Experience (SDX) ensures consistent data security, governance, and control across the data lifecycle and environments and reduces risk and costs. It provides data BWSess policies and status information (metadata, data lineage, metrics, audit trails, and more) as an always-on layer on CDP to meet growing regulatory compliance requirements, reduces business and security risks in handling sensitive data, and provides secure self-service BWSess to data and analytics. With SDX, enterprises set up data BWSess controls and policies once, which are then automatically imprinted across all data and analytics in hybrid or multi-cloud deployments, even as data and workloads move between them.

***100% Open***

The basis of CDP is open-source software (Cloudera Runtime). Licensed under Apache or the AGPL, this approach drives innovation, avoids vendor lock-ins, and delivers enterprise-class software. Using open source ensures that CDP is interoperable with various analytics and business intelligence applications. This interoperability helps reduce the cost and effort required to connect a customer's IT infrastructure to the data and capabilities of delivering CDP.

***Why CDP?***

* **Lower TCO** – CDP offers an integrated data cycle, eliminating the ongoing integration and orchestration development work required to implement business use cases with cloud providers' data or analytics services.
* **More robust security** – CDP is designed to be secure, using multiple capabilities across the data cycle to validate users, authorize actions that users and applications can perform with data, and protect data from unauthorized visibility. Other vendors have some of these security capabilities, and fewer are integrated into all data analytics.
* **Run anywhere** – CDP runs workload analytics across multiple Public Clouds, on-premises Private Clouds, or a combination of each for Hybrid Clouds, with a consistent IT admin and user experience. A single cloud provider can't match this flexibility, critical for data governance, performance, and geographic coverage.

### Cloudera Data Science Workbench

Cloudera Data Science Workbench (CDSW) is an enterprise data science platform that BWSelerates data science and machine learning projects by providing a robust and familiar environment for data science and data engineering with self-service BWSess to business data. CDSW offers a complete solution for data exploration, analysis, visualization, modeling, and application. CDSW makes collaborative, secure, and live data science a reality for enterprises, BWSelerating the delivery of new data products without IT, data science, developers, analysts, and business stakeholders across data science workflows.

***Main Feature:***

* Collaborative hub for enterprise data science with project collaboration, secure collaboration, and simple dependency management.
* Integrated development environment for Python, R, and Scala with support for Spark and connectivity to secure CDH and HDP clusters, and data stored anywhere
* Full cycle model support with experiment tracking and easy model deployment.
* Lightweight job analytics and pipeline management system for data science and data engineering workloads that supports real-time monitoring, sharing of results, and email alerts

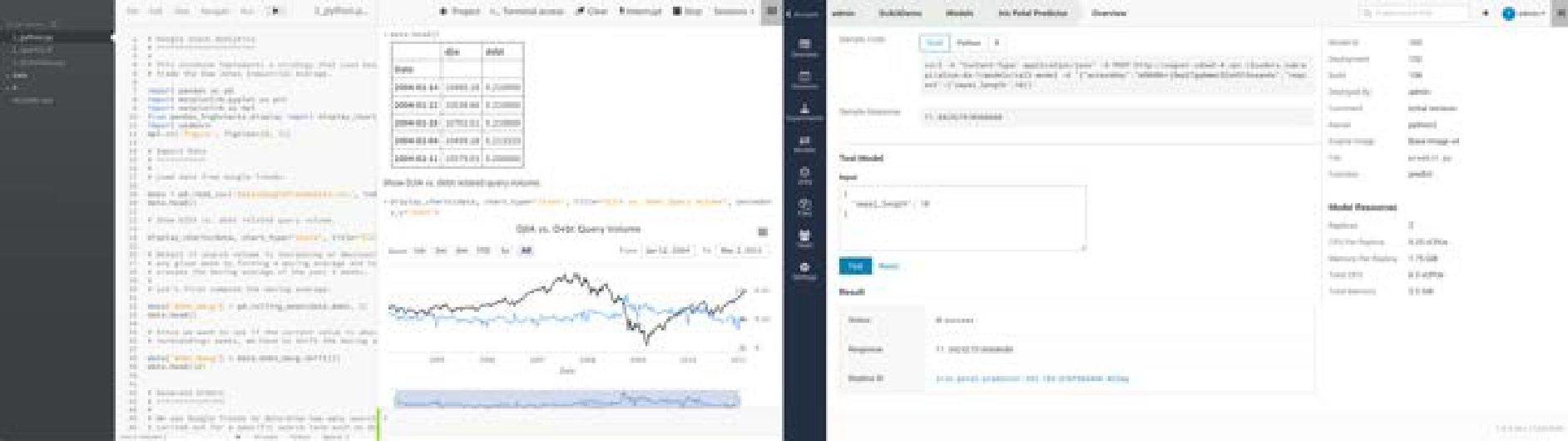


Figure 2 - Cloudera Data Science Workbench

***Deploy model machine learning faster using select development tools***

Data scientists need to bring tools to their data, whether stored on CDH, HDP, or the cloud. Using select editors, including Jupyter Notebooks, RStudio, and more, and multiple languages ​​, including Python, R, or Scala, right from a web browser, CDSW provides data scientists with a rich self-service experience.

Download the latest libraries and frameworks in a customizable project environment. Beyond the Python and R ecosystems, deep learning frameworks such as TensorFlow, PyTorch, Caffe 2, MXNet, DL4J, BigDL, sci-kit-learn, and others evolve. CDSW provides a secure and secure environment to combine the latest open-source innovations with a trust unified platform. Cloudera customers.

A collaborative and shareable project environment ensures diverse data science teams can coordinate towards reproducible standard research and production-ready models and quickly deploy and manage them across stakeholders and end-users.

***Better use of core data management investments***

IT groups sometimes find it difficult to incorporate data scientists into extensive data systems due to their diverse needs, especially where open-source tools are involved. The result is duplication and analytic silos with limited security and governance. Meanwhile, scientists seek to scale work to larger data sets and more sophisticated computing platforms. CDSW helps break down analytics silos and generate value from your enterprise data platform, whether on-premises or in the public cloud.

For modern IT organizations, CDSW can:

* improvements to the existing analytic stack with a collaborative environment that supports the latest open source languages and libraries.
* Better use of core data management investments and introduce new supported tools for data science.
* Expansion of data science beyond closed environments and enabling BWSess to self-service data science

***A complete, secure, and compliant data science solution by default***

IT is responsible for complying with company Directives such as security and governance. This can be not easy when every user BWSesses the environment through a standard interface such as SQL. It becomes even more complicated when each team, user, and project uses different open-source tools. Managing a diverse environment in a secure cluster is nearly impossible. IT, which is set to balance corporate data with data science interests, is often forced to lock down data.

Cloudera helps IT groups and data scientists work together, bringing more users to a shared environment in a way that provides flexibility and compliance. CDSW supports data science work within the regulated and secure CDH and HDP platforms. Users get complete data visibility and governance, auditing, and lineage across platforms.

## Benefit of Cloudera for Data Warehouse

As a Data Platform, Cloudera differs from other Data Warehouse technology in several key highlights as follows:

### Any Data (Structured, Unstructured)

In this digital age, data is the new oil. We sit in figurative gold mine with tremendous data that we have and/or generate. Unfortunately, not all data type can be stored, processed, and analysed simple because the technology was not supported. This happens mostly for unstructured data (type can vary from PDF documents, text, image, sound, to video).

Most data warehouse technology based on traditional RDBMS finds it difficult to store, processed, and analyse unstructured data. Cloudera with Hadoop File System (HDFS) that can be designed as the data lake can store this unstructured data. Later, with services like Spark, this unstructured data can be processed to unearth insight otherwise impossible to obtain. This is why more enterprises uses Cloudera for their Data Warehousing solution.

### Future-Proof

Although Data Warehouse handles batch data processing (true for most RDBMS Data Warehouse tech), some use case requires insights to be processed in real-time or near real-time manners. Future use case in bank organization may require just that. Cloudera can be configured as cluster to handle real-time data processing with the help of Kafka and/or Spark Streaming. Opting Cloudera as Data Warehouse technology provide future-proofing for any data speed need (batch, near real-time, real-time).

### Analytics Ready

If data is the new oil, analytics is the combustion engine. With analytics (specifically data science use cases), bank can venture into advanced use case (like credit scoring with machine learning, cross/up-sell recommendations, churn prediction, customer segmentation and profiling, just to name a few) to gain competitiveness. Cloudera has the analytics capabilities from the getgo. Combined with the capabilities to process and analyzed unstructured data, the insight that can be obtained is limitless. Traditional RDBMS data warehouse technology may require additional software and/or tools to achieve this.

### Cost Effective and Scalable

Cloudera cluster can be scaled-out by adding data nodes when the needs arise. Contrast to traditional RDBMS data warehouse technology that require scale-up by replacing higher spec-ed hardware. This made Cloudera a cost effective implementation.

## Methodology

Described here are the methodology to achieve the proposed solution.

### Requirement Gathering

Requirement Gathering or functional study represents the first step in the implementation stages. During this stage, our consultants coordinate workshops and/or interviews with BWS key Data and/or businss users to identify the data model design for the data warehouse.

Further details on requirement gathering as well as responsibilities of each party as follows:

Table 2 Requirement Gathering Details

| **Key Activities** | **Responsibilities by:** | |
| --- | --- | --- |
| **GIT** | **BWS** |
| Conduct workshops/interview with users to understand and determine data, reports, and dashboards | * Conduct workshops/interview to ensure key business requirements are captured and documented. * Manage writing of meeting minutes (Secretariat work). | * Identify key users * Participate actively in workshops/interview and provide relevant information on a timely basis. |
| Develop and document reporting needs based on information gathered from workshops. | * Gather and document BWS Data Warehouse and reporting/dashboards needs based on pre-agreed format. * Define documentation format and agree with BWS. | * Confirm requirement gathering documentation templates. |
| Review Documentation with BWS. | * Facilitate review of the Business Requirement Document (BRD) with client. * Agree review approach with BWS. | * Confirm review approach. * Conduct detailed review sessions. * Review and confirm accuracy of information on a timely basis. |
| Obtain BWS sign-off for Requirement Gathering documents. | * Facilitate sign-off of Requirement Gathering documentation * Agree sign-off criteria with BWS. | * Review and sign-off Requirement Gathering document on a timely basis. |

*Note:*

* Data and documentation of it (e.g. ERD, data dictionary) for this project is available
* To accelerate Requirement Gathering process, BWS will provide us information related to data such as table name, column name
* BWS will provide us access to source database

### Analysis and Design

As depicted in the previous diagram, system analysis and design begins upon completion of first draft of Requirement Gathering document. During this stage, the technical team begins mapping the requirements captured during Requirement Gathering and developing data model for the data warehouse, aggregation (data mart), as well as the mock-up design of the dashboards.

Sequentially, the business requirements begin to transform, and technical data and information begins to take form. The technical team takes precedence beginning the design works of the system. List of designs involved in the stage are as follows:

* Functional specifications
* Data mapping
* Data warehouse design (Data Model)
* Conceptual (Mock-ups) dashboards design

The deliverables during this stage are:

* Data Model document
* Mock-up document

### Development

After the requirement is completed and the design is finalized (with formal sign-off), development phase can begin.

### Infrastructure Setup

The first step of the development phase is to setup the necessary infrastructures (hardware, software, tools) for development and production environment (production environment can be done later or earlier depends on development environment capabilities):

### Data Integration & Warehouse Development

Data Integration procedures includes Extract, transform, and load (ETL) process that involves:

* [Extracting data](http://en.wikipedia.org/wiki/Data_extraction) from BWS data sources
* [Transforming](http://en.wikipedia.org/wiki/Data_transformation) data to fit dara warehouse data model
* Loading it into the end target which is the data warehouse
  + For first time load, initial loading job will be developed
  + For daily incremental load, incremental loading job will be developed

Assumption for extracting data:

* Data extracts from all identified source systems are available prior to the start of the project.
* Data extract from all data source is clean and does not requires any cleansing process.
* Users and Source system owners/administrators are available to provide clarity / insights on data usages and conventions used by the clients in their respective source systems.

### Dashboard Development

In this phase, the dashboard and/or reports will be developed based on agreed upon mock-ups and connected to the data warehouse and/or data marts to get the query data result.

### Testing

Every complete product will be tested before it’s going to production environment. For BWS Data Warehouse & Dashboard, User Acceptance Testing (UAT) will be conducted after the development is completed. Upon sign-off of the tests (UAT), the system will be handed to the project owner.

### User Acceptance Testing (UAT)

*User Acceptance Testing (UAT)* is a process to obtain confirmation by a Subject Matter Expert (SME), preferably the owner or users of the object under test, through trial or review that a system meets mutually agreed upon requirements. UAT is the final phases before the system Go-Live.

This approach has been proposed in order to achieve the following goals:

* To meet the objective or scope of the actual UAT that will be carried out
* To make sure all critical processes are not to be missed out
* To familiarize users with the necessary tools
* To meet the milestone of the Data Warehousing development project

### Preparing the Test Plan

The system will be fully tested by the client against the requirements defined in the Design stage.

Looking at the UAT from a high level, these are the basic steps that need to be undertaken:

Table 3 Basic steps for UAT

|  |  |
| --- | --- |
| **Step** | **Description** |
| **Test Strategy** | Decide the approach of the testing in terms of people, tools, procedures and support. This entire document is all about the Test Strategy. |
| **Training** | Training for End-Users before preparation of Test Scenarios and Test Scripts. |
| **Test Scenarios** | The situations to be tested out. |
| **Test Scripts** | The actual inputs that will be used, normally in a step-by-step format. The expected results are compared against the actual results. Prepared by the Users. |

The constraints of test exercises are stipulated below:

* Users have to conduct the UAT during working hours with no major interruption of their daily work.
* Limited time of development and unit testing.
* Any interruption from 3rd party on the UAT environment will affect the timeline.

# Architecture

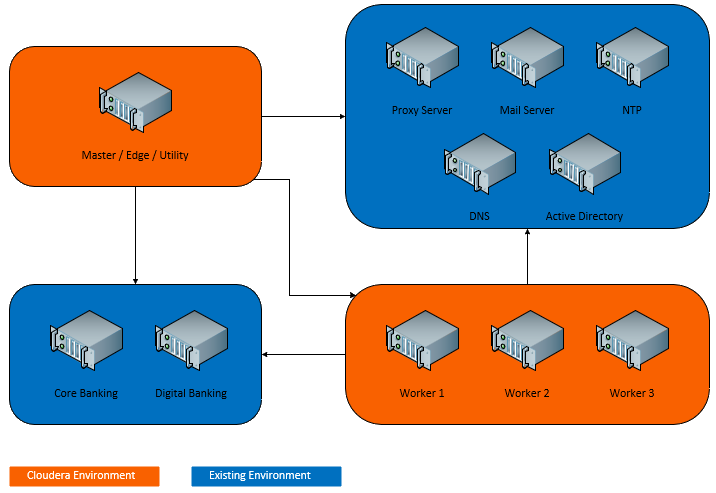
## High-Level Solution Architecture

Graphical user interface, timeline

Description automatically generated with medium confidenceThe following is the high-level solution architecture for the data warehouse with Cloudera.

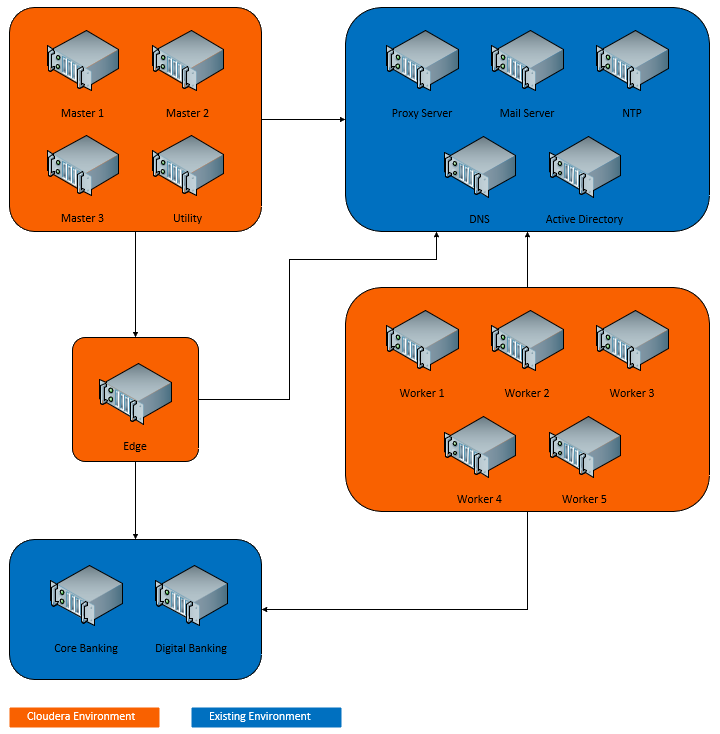
1. The data source will be ingested from Oracle, DB2, and PostgreSQL using Nifi (ETL/ELT tools).
2. The ingested data will then be stored on the data lake (HDFS)
3. data will be transformed based on the data model design into facts and dimensions to Hive tables
4. Data mart will then be built for the necessary aggregation model from the data warehouse.
5. Dashboard will be developed using Superset (can be replaced with other BI tools) that connected to data warehouse and data mart)

## Development Environment

The following is proposed development environment topoology

## Production Environment

The following is proposed production environment topoology



# Project Management

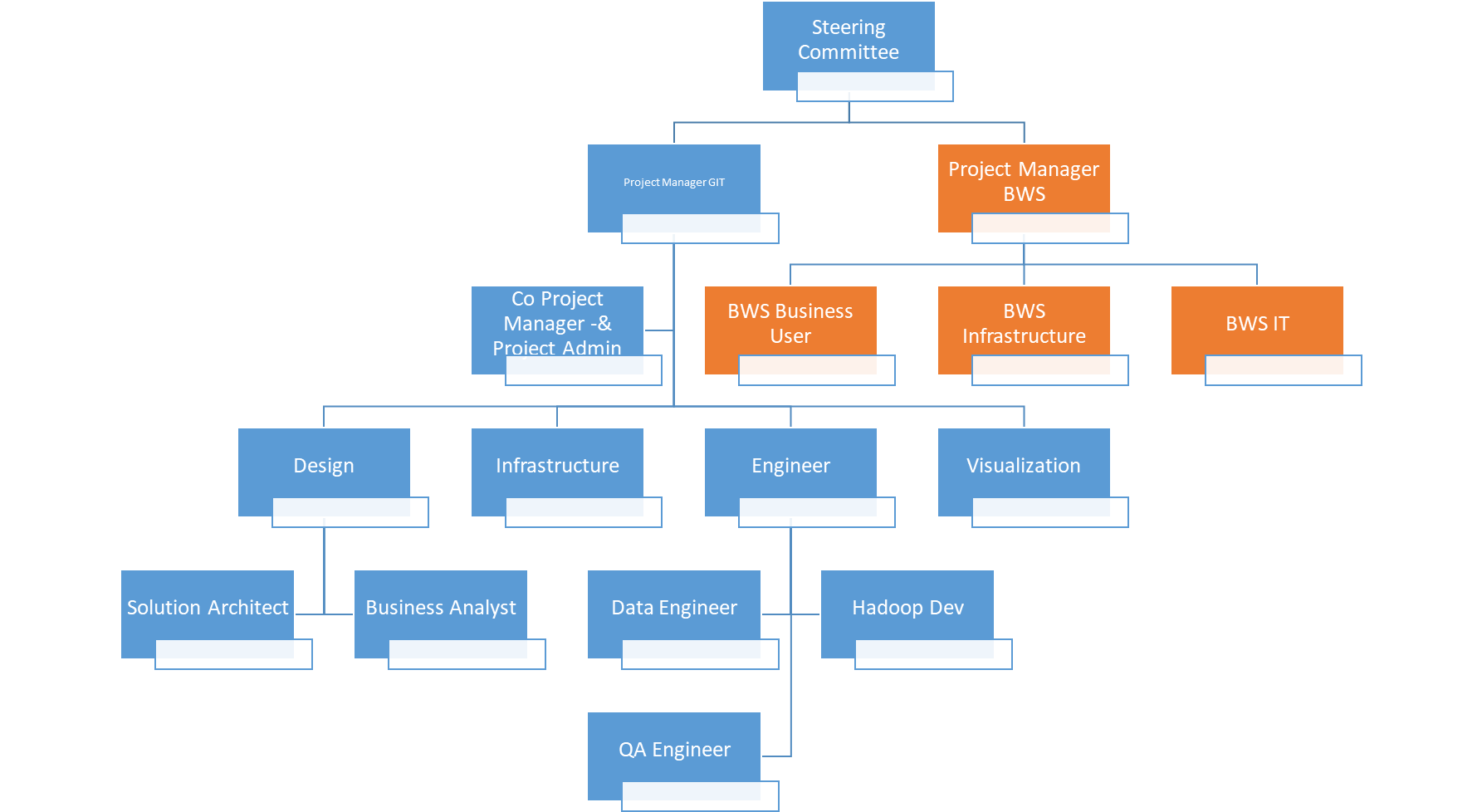
## Timeline

The following is the timeline and Work-Breakdown Structured proposed for this project.



## Project Organization Structure

The following is the Project Organization Structure proposed for this project.

******

## Key Success Factor

The following is the key success factor for this project:

* Full commitment from BWS project team throughout the project, to work with GIT consultants to determine the suitable scope, assumptions, data structure, reports structure and any applied rule sets in this project
* Availability of the required files from source systems, i.e. extracts of the incremental (delta) records for on-going load from production system, the manual data must be on csv format (if any)
* Strong sponsorship from BWS Senior Management
* Clear ownership of Data throughout BWS as an enterprise, throughout the project as well as after the project is completed, to continuously review and enhance the utilization of this new data warehouse system to support operational and business

## Assumptions

The following is the key success factor for this project:

* All the Hardware and Data sources needed will be obtained and available to be used when conducting the project. However, there might be a lack of information about tables and data sources required for the project life cycle. This can hinder project management and implementation efficiency because not everything needed is available. (currently we do assumptions on the table size relevant to the project. Newly identified information may result to additional project timeline and/or cost)
* All the people that are involved at the beginning of the project will remain and stay in the project until the project is completed. However, it is crucial to take note that there may be workers that may withdraw from the project throughout the project life cycle.
* All the resources (Data Source, Data dictionary, List of Table) will be available to be used throughout the project life cycle. However, there might be issues where the resources are not available or not complete to the project's progress.
* Members of the project team would have access to the system they need to execute their respective tasks on schedule, including all support things during working time. However, access may have limitations for everyone. about limited access need assistance from the internal team.
* All relevant stakeholders will come to the next meeting as scheduled. However, there might be instances in which the stakeholders are unavailable to attend.
* The expected project’s timeline can be met, and the project will complete within the expected time.
* The scope and specifications of the project will not undergo changes when the project takes place. However, when conducting the project, there might be cases where the scope and specifications need to be altered to cater to the requirements and needs of the project.
* Maintain good communication and keep track of the Timeline with the Team Project and Internal Team during weekly meetings or the specific meeting.
* The IT support for the project will be able to cater to technical difficulties such as system errors, network congestion, connection problem, and technical errors.
* There are minimal network congestion and strong internet connection in the location where the project is conducted. However, sometimes, connection disruptions might occur due to other conditions that may be related to the environment or surroundings of the project.