**CEB 1250 - Big Data Storage**

**Project Information Database and Data Warehouse Design**

**Ferdinand Widjoj Putra**

**John Rinon**

**Michael Minkiewicz**

Table of Contents

[0](#_Toc11270819)

[1. Overview 3](#_Toc11270820)

[**1.1.1** Customer 3](#_Toc11270821)

[1.2 The scope of the project: 4](#_Toc11270822)

[2. Assumptions/Constraints/Risks 5](#_Toc11270823)

[2.1 Assumptions 5](#_Toc11270824)

[**2.1.1** Software & Hardware Environment 5](#_Toc11270825)

[**2.1.2** End-User Characteristics 5](#_Toc11270826)

[**2.1.3** Operating Systems: 6](#_Toc11270827)

[2.2 Constraints 6](#_Toc11270828)

[**2.2.1** Project Timeframe 6](#_Toc11270829)

[2.3 Risks 7](#_Toc11270830)

[3. Design Decisions 8](#_Toc11270831)

[3.1 Key Factors Influencing Design 8](#_Toc11270832)

[**3.1.1** General Factors 8](#_Toc11270833)

[3.2 Functional Design Decisions 9](#_Toc11270834)

[3.3 Database Management System Decisions 12](#_Toc11270835)

[3.4 Performance and Maintenance Design Decisions 13](#_Toc11270836)

[**3.4.1** Database Setup 13](#_Toc11270837)

[**3.4.2** Database Design for archiving 14](#_Toc11270838)

[4. Detailed Database Design 15](#_Toc11270839)

[4.1 Project Information Database Design 15](#_Toc11270840)

[**4.1.1** Conceptual Data Model 15](#_Toc11270841)

[**4.1.2** Logical Data Model 16](#_Toc11270842)

[**4.1.3** Physical Data Model 17](#_Toc11270843)

[**4.1.4** Data Dictionary 18](#_Toc11270844)

[**4.1.5** Expected Data Volumes 20](#_Toc11270845)

[**4.1.6** Life Expectancy 20](#_Toc11270846)

[4.2 Business Intelligence Environment 21](#_Toc11270847)

[**4.2.1** Data Warehouse Architecture 21](#_Toc11270848)

[**4.2.2** Data Warehouse Hierarchy 21](#_Toc11270849)

[**4.2.3** Data Granularity 22](#_Toc11270850)

[**4.2.4** Data Warehouse Multidimensional OLAP Model 23](#_Toc11270851)

[**4.2.5** Data Warehouse Schema 24](#_Toc11270852)

[4.3 Performance Monitoring and Database Efficiency 25](#_Toc11270853)

[**4.3.1** Operational Implications 25](#_Toc11270854)

[**4.3.2** Data Transfer Requirements 25](#_Toc11270855)

[**4.3.3** Data Formats 25](#_Toc11270856)

# Overview

Service Excellence corporation is a mid-sized company providing engineering services primarily to the aerospace industry. Headquartered in the United States, Service Excellence Corporation employs approximately 36,000 staff and services 9000 clients across North America, Europe and Asia.

In Canada there are 15 locations across the country employing 4500 people, managing 2000 engagements per year.

Each engagement represents a defined scope of work, executed over a finite period of time with a fixed budget assigned by resource type, by month. Information related to the project budget is typically captured on an ad hoc basis on the local computers of the assigned project manager.

As part of a Project Execution Optimization (PEO) initiative, senior management of the Canadian operations has issued a directive to improve the organizations ability to manage and have visibility on actual project costs vs. budget, & maximize employee utilization/allocation between projects this will allow to maximize resources utilization and profits.

As part of this initiative a mandate has been given to develop a data storage system that will:

1. Consolidate all information related to project budgets into a single sources of truth
2. The system and data must be accessible to employees anywhere via their laptop or desktop equipment.
3. Facilitates the ability to view budget vs. actual costs by:
   1. Project overall
   2. Project by month
   3. Project by resource type

### Customer

1. Easily track employee utilization and resources for optimal allocation
2. View must be available on a weekly, monthly, quarterly and annual basis.
3. The system should be scalable to allow other regions of the organization to be able to adopt the system should a wider deployment be deemed beneficial.
4. The system should minimize the amount of redundant information
5. The system should, to the greatest extent possible, leverage existing platforms used by the company.

The current enterprise system used by the organization is SAP. It already contains all information related to:

1. Employees and the resource type they belong to.
2. Hours entered by employees and the project they charge to.
3. Purchase orders for goods or service and the project they are related to.

## The scope of the project:

The objective of this project is to optimally store and access the data to facilitate decision making. Since there data coming from different data sources, data consolidation will be the main challenge in this project. Hence, the teams propose solution is to:

Develop an ***Operational Database ( Project Execution Database (PED))*** that will serve as the consolidate source of all project related data for the organization.

Establish a ***Business Intelligence Environment*** that will include a data warehouse infrastructure that will pull data from the established enterprise databases and the PEO database to extract, transform and load the data to facilitate analysis providing insight/recommendation to the executive team.

# Assumptions/Constraints/Risks

## Assumptions

### Software & Hardware Environment

#### IT Facility Requirements & Servers

Space exists in current IT facilities to house the infrastructure necessary to establish both the operation database and BI environment.

#### IT Maintenance Tasks

All maintenance required to support the infrastructure will be provided by existing teams

#### Access Control

Access, login and read/write credentials will be managed by existing enterprise servers.

#### Data Architecture

HDFS

#### Apache Spark

Analytics engine currently utilized at Service Excellence

#### Microsoft Office 365 Package

Employees have access to Microsoft applications & visualizations tools.

#### Relational Database Environment

Service Excellence uses relational Database throughout the entity for the database environment.

#### SAP

Enterprise software in place to manage business operations

#### Confluence (WIKI)

Entity uses Confluence to document the different projects in order to ensure replication in case of loss of key personnel or corrupt data.

### End-User Characteristics

#### Users (Resources) skill base

Project manager: Proficient in MS Office tools set including power BI

#### BI Team

Proficient in Python and SQL

#### Data Pipeline (end-user)

The data input pipeline for all database has been properly establish with responsive web application

#### Database Permission

Database permission has been properly set up with 3 levels of access which are user(read), admin/developer(edit)

#### Normalized Database

Relational database has been properly normalized to 3rd level of normalization.

#### Data Naming Standard

Data model adhere to a solid naming standard, it will also contain definitions on its tables, columns, relationships, and even default and check constraints, so that it is clear to everyone how they are intended to be used.

### Operating Systems:

#### Microsoft Windows 10

All employees use Microsoft Windows 10 as operating system.

## Constraints

### Project Timeframe

Short timeframe to deploy, this project has been given 6 month timeframe to develop, test and deploy the system.

#### Budget

Limited funds are available the allocated budget for this project is $20,000 USD.

#### IS/IT Capacity

The IT department is primarily structured to support off-the-shelf software and hardware issues. There is a small team that is responsible for a handful of in-house applications. They have limited capacity to take on a major new technology deployment.

A process and tools team is available to assist in the development and documentation of the process as well as the training activities that would be needed to deploy the solution.

#### Technology

The new operational database (Project Execution Database (PED)) must be fully compatible with Service Excellence’s current software, hardware, & operating systems already in place/

Therefore based on the budget, timeframe, IS/IT Capacity constraints, Service Excellence must develop the operational database with the technologies already being utilized internally.

## Risks

#### Data Backups

All data is backed up daily to ensure its availability in case lost of data

#### Cyber Security

To ensure data security, access has been properly controlled and can only be accessed through VPN.

#### Different Data Sources

Data ingestion from different sources therefore risk of lost or corrupt data, this is why data will be backed up on daily basis.

#### Data Leaks

To mitigate this risk Service Excellence’s administrators use SSL- OR TLS encrypted communication platforms.

#### Segregation of duties

Need to ensure that for the employees who are working on this project not everyone has administrator user powers, there needs to be different levels of access this will limit the power of user accounts and therefore limit the risk of a hacker taking control of the database.

#### Deployment Failures

When the database is deployed it does do what is expected of it, to mitigate this risk the project will be deployed through a DEVOPS approach using agile method.

# Design Decisions

## Key Factors Influencing Design

Physical database design is key to guarantees good performance.

### General Factors

#### System Criticality

This system should be considered as critical to the operation of the business. As such all backup and disaster recovery protocols to ensure the integrity and safety of the data must be in place once this system goes live.

#### Primary Users Requirements

The primary users of the system will consist of the project managers and the business intelligence group.

The project managers will be using the system to enter the project related data and generate the key performance indicators (KPI) related to the performance of their project. As such the following requirements apply:

#### User Interface

The interface used to enter the project related data must be:

1. Simple to use
2. Minimize the need to enter redundant data
3. Perform checks to prevent the creation of the same data in different ways (i.e CAE Inc, CAE, CAE Incorporated)
4. KPIs and reports, once configured, must be able to be saved and re-run on an ad hoc or scheduled basis.

#### Accessibility

Access to the PED must be available from laptops or desktops on the company’s internal network or via the companies VPN technology when using external network.

#### System Resources

The project management team in Canada fluctuates between 150 to 300. The Business Intelligence group consists of a team of 5 people. Thus the PED shall be capable of handling 100 concurrent users.

Each project should be allocated approximately 5GB of storage space to house the project related data.

Technology should have the flexibility to be scale up should other regions begin deploying the system as part of their operations.

#### Performance

Performance is not the most important design factor, since decisions don't need to be made in real time by the users. The data granularity of the data used by users does not change dramatically from day to day, and decisions need to be made or tracked on a weekly or monthly basis.

#### Availability

Users do not need the information in real time, however the information needs to be available at all times for the users to be able to make business decisions, since we are dealing with tracking of resources over the length of a project (projects can span over many months or years)

#### Security

The information will be coming from projects that are located all around the world. The data contains sensitive employee information (such as name, hourly rate, etc). Additionally, the data contains sensitive information concerning all the entities projects and their clients. It is very important that the data is transferred in a secure manner.

#### Privacy

## Functional Design Decisions

#### Data retrieval

Easy retrieval of information based on query and transaction frequency.

#### Nomenclature consistency

Nomenclature of entities and attributes must be consistent in all cases to facilitate data retrieval.

Ability to retrieve historical dataFor analytical purpose, historical data retrieval needs to have an independent query.

#### Able to retrieve/download aggregated data

Another functionality for reporting & analytical purpose.

#### Possibility to modify database content

Data Manipulation Language (DML) is required to properly manipulate the data.

#### Possibility to modify structure of the database

Data Definition Language (DDL) is required if database structure requires modification.

#### Able to administer user rights

Data Control Language (DCL) is needed to grant/revoke access to designated users.

Able to modify previously entered record in database

Existing records have to be editable by designated user.

As described in the overview section, the primary users of the data through the operational database that will feed the data warehouse will consist of the project managers, finance/accounting group, HR group & the business intelligence team that will be supporting the data pipeline.

The project managers will want to gain insights on the following information:

* 1. Consolidate all information related to project budgets into a single source of truth
  2. All projects in the pipeline including the projected completion date with the following levels of granularity:
     1. Consolidated view of all projects
     2. Projects per region
     3. Projects per year
     4. Projects per quarter
     5. Projects per month
  3. Facilitates the ability to view budget information vs. actual costs at the following levels of granularity:

1. Project performance as an overall
2. Project performance by month
3. Project by resource type
4. Project performance per customer
   1. Track employee utilization and resources for optimal allocation between projects, the level of granularity for this information will be the following:
5. employee utilization per week
6. employee utilization per month
7. employee utilization per quarter
8. employee utilization per year
   1. All information view must be available on a weekly, monthly, quarterly and annual basis.

The Accounting/Finance group will want to gain insights on the following information:

* 1. View budget information vs. actual costs at the following levels of granularity:

1. Project performance as an overall
2. Project performance by month
3. Project by resource type
4. Project performance per customer
5. Global performance per region
6. Global performance as a whole
   1. View project profitability at following levels of granularity:
7. Overall profitability
8. Profitability per project
9. Profitability per region
10. Profitability per customer
    1. All information view must be available on a weekly, monthly, quarterly and annual basis.

The HR group will want to gain insights on the following information:

1. Will want to track resource capacity in order to determine if entity is over or understaffed at a given time:
   1. Overall resource capacity
   2. Resource capacity per region
   3. resource capacity per project
2. All information view must be available on a weekly, monthly, quarterly and annual basis.

The end users will be able to access the data directly from the data warehouse the data files will be compatible with the windows operating system that all end users have access to. The data extracts will be compatible with Microsoft Office 365 software (software for which every employee has a license). The users will be able to create dashboard threw PowerBI to visualization the different insights listed above through an interactive real time dashboard. Additionally the different users will be able to extract CSV files to use as reports and extracts based on different ad hoc needs.

The data will be updated in batch on a daily basis so that the users PowerBI dashboards will be updated to reflect the updated information.

The Business Intelligence employees will work with the project managers, accounting/finance team & HR team to develop the PowerBI dashboards based on the users needs. Therefore the project managers, accounting/finance team & HR team will only have a read access to the dashboard. Since they will be using the information as insights to make business decisions.

The Business Intelligence (BI) team will have administrator access to the dashboards since they will be the ones to make modifications to dashboards based on the specifications provided by the business units. The BI team and the business units will work in agile method to make modifications to the dashboard based on the evolving needs of the business units.

The specifications and audit trail for the dashboards will be documented in Confluence in order to ensure proper replication of the dashboards in case of loss of key personnel or PowerBI becomes corrupt.

The decisions described above were made based on compatibility with the current software available to end user employees. Therefore there will be little training and learning curve required from the end users, since the data will be visualized on applications that the end users are already using. These decisions are to ensure that the project is rolled in short time frame to respect the timeframe and budget.

## Database Management System Decisions

Operational data source came from multiple sources. Apache Kafka will be used as pipeline to channel data into data lakes. Kafka is chosen as the company will be moving towards real-time streaming data architectures to provide real-time analytics. Other reasons include fast, scalable, durable, fault-tolerant and most importantly operational simplicity.

All data will be channeled to Hadoop as the data lake with following reasons:

#### Range of data source

The data collected from various sources will be of structured or unstructured form. A lot of time would need to be allotted in order to convert all the collected data into a single format. Hadoop saves this time as it can derive valuable data from any form of data.2.

#### Cost Effectivity

In conventional method, company spent considerable amount of their profit into data storage. In certain cases they even had to delete large sets of raw data in order to make space for new data which allows possibilities of losing valuable information. By using Hadoop, this problem was completely solved. It is a cost-effective solution for data storage purposes. This helps in the long run because it stores the entire raw data generated by a company

#### Speed

Every organization uses a platform to get the work done at a faster rate. Hadoop enables the company to do just that with its data storage needs. It uses a storage system wherein the data is stored on a distributed file system. Since the tools used for the processing of data are located on same servers as the data, the processing operation is also carried out at a faster rate.

#### Multiple Copies

Hadoop automatically duplicates the data that is stored in it and creates multiple copies. This is done to ensure that in case there is a failure, data is not lost.

Using Kafka & Hadoop in the system will ensure integrity and performance of the system. Data will be then pulled to Data Warehouse for different needs such as Business Analysis or to develop Machine Learning algorithm. In case the end user accidentally lose data in the data warehouse, the team can simply pull the data again from Hadoop since it stores all original datasets.

The DBMS will consist of MySQL installed within WAMP package for Apache web server.

The main reasons for this are as follows.

#### Speed

For the scope of the project, the speed would be more than enough for all required database functionality. Speed is not the main priority for our users as described in section 3.1 users don’t need to make real time decisions.

#### Ease of use

High performance system but relatively simple database system to use and a lot simpler to set-up and administer than larger system.

#### Query Language Support

Use standard Query language for modern database system.

#### Capability

It’s multithreaded system. Hence, many instances can connect to it at the same time. It can be accessible interactively using several interfaces that allow queries and view the results: command line, web browsers(i.e. phpmyadmin), Git Bash. Programming interfaces are available for many programming language such as Python, PHP, Ruby, etc. ODBC which is database communication protocol can be used for quick UI prototyping using MS Access. Therefore is compatible with the hardware and software currently being used at the company

#### Connectivity and Security

Fully networked. The database can be accessed from anywhere simply by knowing general info such as IP address, Port, credential and database name. MySQL has strong access control which can prevent unwanted access from different user.

#### Availability and Cost

It’s an open source system. This means that MySQL is available without cost for most in-house uses, therefore fits the budget constraints allowed for this project.

#### Open Distribution and Source Code

For development, there are many users that have worked using MySQL. Many bugs-fix and algorithm are readily available to be used. Employees and IT workers are familiar with MySQL therefore no additional training requirements or learning curve for the employees.

## Performance and Maintenance Design Decisions

### Database Setup

On corporate server, the team will set up a windows Virtual Machine(VM). The VM will prevent corporate service interruption should the Operating System(OS) need to be rebooted. Within the VM, WAMP package will be installed. Apache web server will be used in port 80 and port 3306 for MySQL (default port setup for the system). Depending on availability, port may be modified as needed. DBMS will be designed and developed using web-based GUI (phpmyadmin) or directly through command line.

### Database Design for archiving

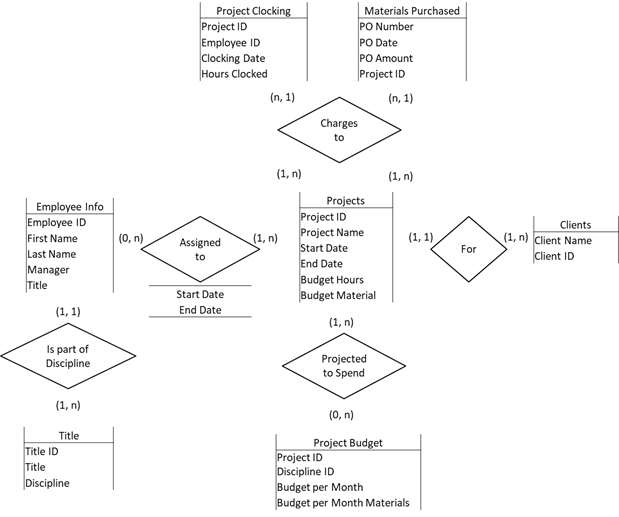
HDD for closed projects and employee data SSD for current projects monthly backups stored off site

# Detailed Database Design

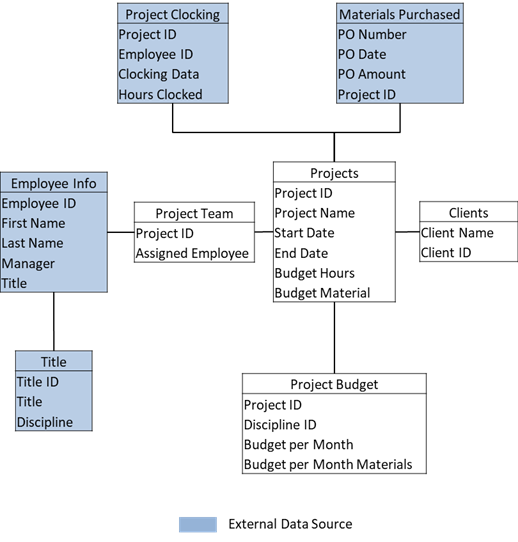
## Project Information Database Design

This section will outline the tools used to develop the operational database that will implemented to address the requirements related to entry and management of project information. Section 4.2 will focus on the tools used to develop the business intelligence environment.

### Conceptual Data Model

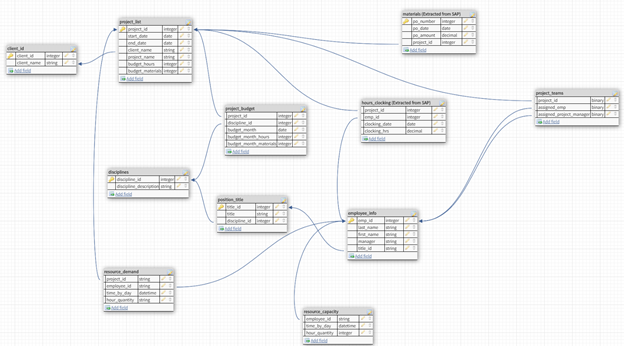
**

### Logical Data Model

**

The external data source identified in the logical model is data that will pulled from the enterprise instance of SAP which is the main repository for this information. They are depicted in this logical model to define the relationship between the PEO database and SAP. Thus the only entities that will specifically be part of this database design are those in white.

### Physical Data Model

**

### Data Dictionary

| ***Data Element Name*** | ***Type*** | ***Length*** | ***Source*** | ***Constraints*** | ***Validation Rules*** | ***Audit and data masking*** |
| --- | --- | --- | --- | --- | --- | --- |
| *client\_id* | *Integrer* | *N/A* | *UI* | *System generated* |  |  |
| *client\_name* | *String* | *50* | *UI* | *Entry of this data on the client\_id table will be limited to superusers* | *Cannot already exist and*  *prompts user if similar client\_names already exist* | *Selection of the "client\_name" on the project\_list table will constrained to the list of clients defined in the client\_id table* |
| *Project\_id* | *Integrer* | *N/A* | *UI* | *System generated* |  |  |
| *start\_date* | *Date* | *N/A* | *UI* |  | *At entry must be greater than current\_date* | *Will use drop down caledar functionality to eliminate data entry errors* |
| *end\_date* | *Date* | *N/A* | *UI* |  | *Must be greater than start\_date* | *Will use drop down caledar functionality to eliminate data entry errors* |
| *project\_name* | *String* | *50* | *UI* | *No special characters* |  |  |
| *budget\_hours* | *Integrer* | *N/A* | *UI* |  |  |  |
| *budget\_materials* | *Integrer* | *N/A* | *UI* |  |  |  |
| *budget\_month* | *Integrer* | *N/A* | *UI* |  |  |  |
| *budget\_month\_hours* | *Integrer* | *N/A* | *UI* |  |  |  |
| *budget\_month\_materials* | *Integrer* | *N/A* | *UI* |  |  |  |
| *project\_id* | *Integrer* | *N/A* | *UI* | *System generated* |  |  |
| *assigned\_emp* | *String* | *50* | *UI* | *must exist in SAP* | *validates against SAP employee list* | *will use drop down containing employees configured in SAP* |
| *dsicipline\_id* | *Integrer* | *N/A* | *External (SAP)* | *AS per SAP configuration* | | |
| *discipline\_description* | *String* | *255* | *External (SAP)* |
| *title\_id* | *Integrer* | *N/A* | *External (SAP)* |
| *title* | *String* | *50* | *External (SAP)* |
| *emp\_id* | *Integrer* | *N/A* | *External (SAP)* |
| *last\_name* | *String* | *50* | *External (SAP)* |
| *first\_name* | *String* | *50* | *External (SAP)* |
| *manager* | *String* | *100* | *External (SAP)* |
| *po\_number* | *Integrer* | *N/A* | *External (SAP)* |
| *po\_date* | *Date* | *N/A* | *External (SAP)* |
| *po\_amount* | *Decimal* | *N/A* | *External (SAP)* |
| *clocking\_date* | *Date* | *N/A* | *External (SAP)* |
| *clocking\_hrs* | *Decimal* | *N/A* | *External (SAP)* |

### Expected Data Volumes

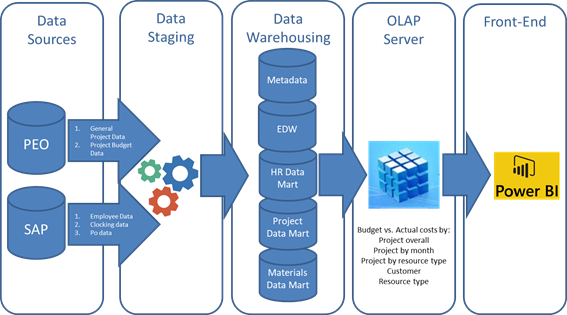
As mentioned on section 3, all data came from multiple sources and backed up several times a day for security purpose. Most data generated are mainly strings, integers and datetime. Hence, the total volume of data range depending on the amount of transaction which will be performed (1-2TB per day including backups).

### Life Expectancy

Data analysis will be done with frequency as mentioned above. Hence, keeping historical data is crucial to be able to identify pattern and provide insight to executive team. As per regulation, the company is required to keep Operational Data Stored for 7 years.

## Business Intelligence Environment

### Data Warehouse Architecture

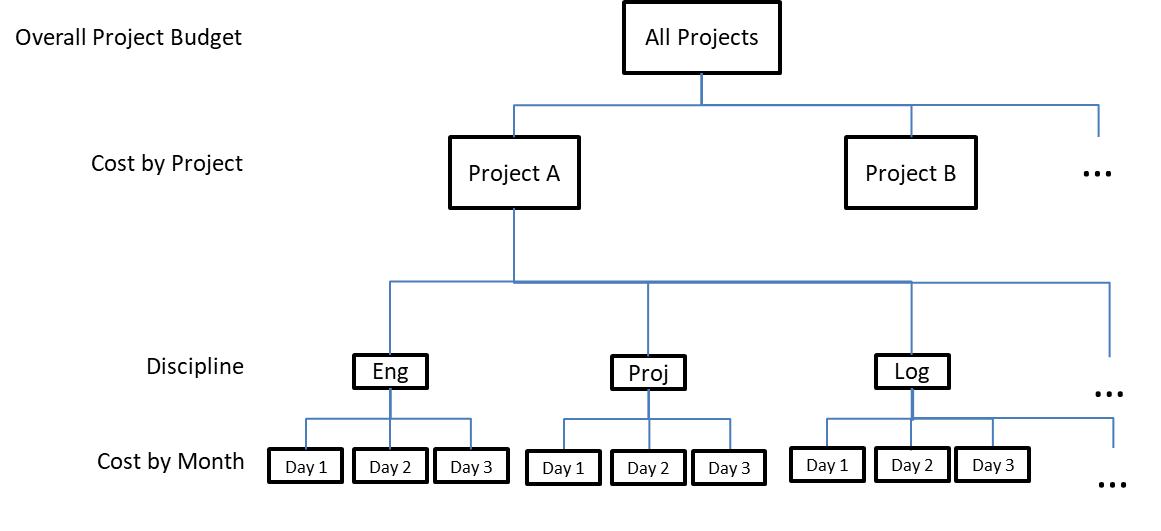
**

### Data Warehouse Hierarchy

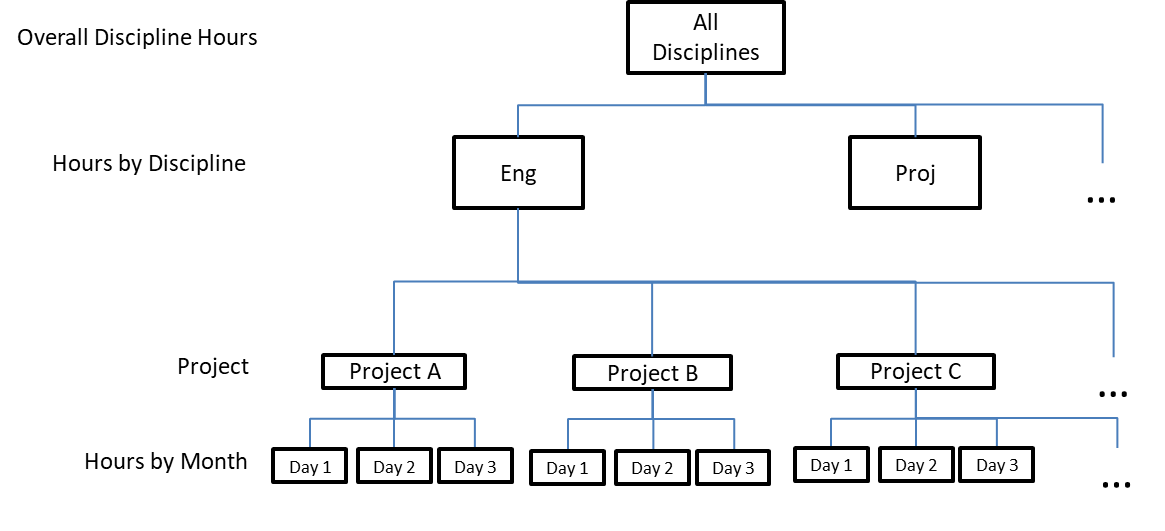
There are two questions we will be trying to answer using our data warehouse capabilities

1. How much are we spending per project? Secondary questions are:
   1. Are we within our allocated budget by discipline?
   2. Are we within out budget by Month?
2. How many hours are we applying to projects? Secondary questions are
   1. How many hours are we spending by discipline?
   2. How many hours are we spending by project?

In order to answer these questions we developed two data warehouse Hierarchies that will allow us to do this

Project Cost Data Warehouse Hierarchy

**Project Cost Data Warehouse Hierarchy**

. **

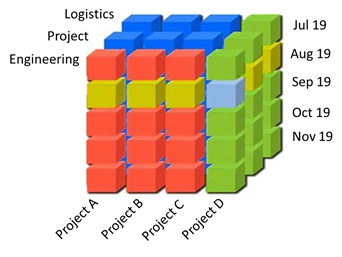
**Manpower Utilization Data Warehouse Hierarchy**

### Data Granularity

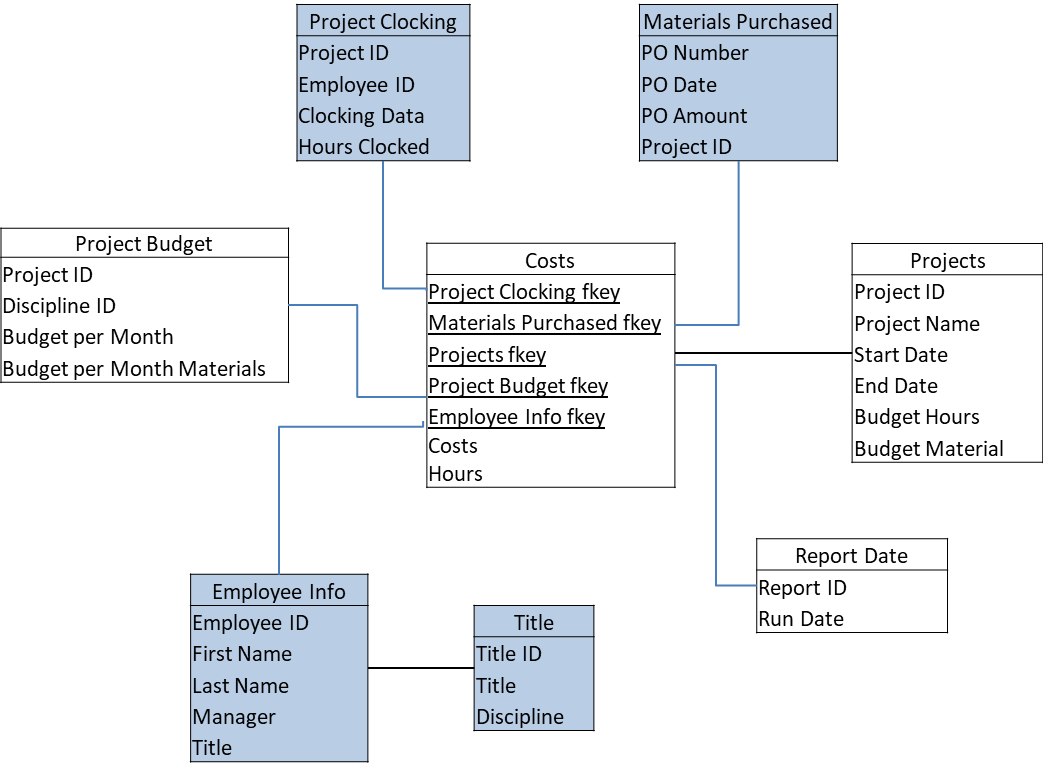
Although cost data is available by hour we decided that level of granularity did not add value to the analysis we wanted to do so we identified day as the lowest level of granularity for the Project cost data warehouse hierarchy.

Similarly there are several lower levels of granularity in the Engineering Discipline Dimension however we decided consolidate all hours charged by engineering resources into a single Engineering bucket for the Hours data warehouse hierarchy.

### Data Warehouse Multidimensional OLAP Model

**

### Data Warehouse Schema



## Performance Monitoring and Database Efficiency

### Operational Implications

Data Transfer will be performed with predefined frequency mentioned above and additional ad-hoc request when required. Data will be refreshed & updated as soon as there is a new input to ensure real-time data integrity.

System backups and restoration will be handled over the weekend with dedicated team to ensure operational readiness during weekdays.

### Data Transfer Requirements

Main requirements for data transfers are speed and reliability. It is important to be able to move large digital data sets quickly across global natively supports direct-to-cloud object storage transfers to and from on-premises infrastructure

Data transfer also requires to be flexible meaning that it should be able to run on commodity hardware, so you can use your existing infrastructure or leverage any other infrastructure or storage type, regardless of location

Lastly, the data transfer has to be secured since it contains company’s confidential data by using open standards cryptography for user authentication, data encryption and data integrity verification

### Data Formats

Data formats for both sending and received system will be handled in Kafka and Hadoop. Since data came in from multiple sources, all data format will be standardized to ensure there won’t be any problem on the receiving end. This will ensure data consistency and integrity for end users.