Project Scope

Project Title: Dataware Housing Solution	**Project Number:** 12
Technologies : Elastic search, Cassandra, Hadoop, SQL	**Anticipated Project Start Date:** 28th October, 2015
Student Mentor : Ashish Kumar	**Date Prepared:** Thursday, 29th October, 2015
Faculty Mentor : Dr. Vasudev Varma	**Estimated Completion Date:** 21th November, 2015

Team Members:

- Pulkit Aggarwal
- Ankur Shrivastava
- Anurag Tyagi
- Tushan Jain

Purpose of Project:

Data Warehousing solution

- Building a data warehousing solution based on Hadoop and Spark SQL integration.

Background:

Data warehousing solutions are required because:

- A Data Warehouse Enhances Data Quality and Consistency
- A Data Warehouse Provides Historical Intelligence
- A Data Warehouse Generates a High ROI
- A Data Warehouse Saves Time
- A Data Warehouse Delivers Enhanced Business Intelligence

Introduction :-

Project mainly focus on the ability to review historical trends and monitor near real-time operational data. Since information stored in a data warehouse is critical to organizations for decision-making and predictive analysis there must be a tool which can manage whole data efficiently.

Technologies to be use :-

a. Elasticsearch :-

Elasticsearch is an open source search and analytics engine (based on Lucene) designed to operate in real time. It was designed to be used in distributed environments by providing flexibility and scalability.ElasticSearch offers ways to extend searching capabilities through the use of APIs and query DSLs. Since ElasticSearch is a document-oriented search engine. Each record in ElasticSearch is a structured JSON document which enable to process query fastly and reduces long waiting time.

b. Apache Cassandra :-

Cassandra follows a peer-to-peer architecture, instead of master-slave architecture. Hence, there is no single point of failure in Cassandra which makes our system fault tolerance. Also it provide elastic scalability which helps us to expand our database as much as possible. Also Cassandra can be deployed across multiple, geographically dispersed data centers which help us to extend our project if time permits.