**Dissertation Title**

**Data Warehouse on Cloud Enabled Platform**

### BITS ZG629T: Dissertation

by

Debasmita Nandi

ID No.

# Dissertation work carried out at

## Infosys Ltd, Bhubaneswar, Orrissa,India

Submitted in partial fulfillment of M.S. Software Systems / Manufacturing Management degree programme

Under the Supervision of

Name and Designation of Supervisor,

Infosys Ltd,Bhubaneswar,Orrissa,India

****

**BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE**

**PILANI (RAJASTHAN)**

**July, 2012**

1. Background (Relevance of the project to the current work environment in the employing organization)

Currently working on Advanced Web Applications with large scale Database.

The Project work will help Employing Organization to adopt Cloud Technology and implement fast Analytics using cutting-edge database.

1. Objectives

As a programmer it’s amazing and highly motivating to experience the recent path-breaking innovations in the fields of Cloud Computing which is being adopted by all small, medium, large-scale Enterprises.

Cloud Enabled Integration Platform (CEIP) and Cloud Enabled Application Platform (CEAP) are emerging as the best cloud offspring. CEIP integrates various services (both SaaS and on-premise) on cloud.

CEAP offers rapid application development on cloud. In both the cases users do not need to worry about software and hardware infrastructure as all CEIP and CEAP vendors already depend on (homegrown or third-party) stack of cloud technologies (IaaS, PaaS and SaaS [in some cases] ). This emerging software model opens up the possibilities of implementing many different interesting services and solutions by leveraging Cloud-ready Scalable technologies in all the domains of Database, Messaging, Application Server, Event Processor etc.

One such promising Cloud-enabled Software solution is ‘Data Warehouse Service on Cloud’. This is an open ground with full of possibilities and innovations.

Our objective is to build a Cloud-based Data Warehouse solution for analyzing heterogeneous set of data for diverse domains by leveraging the Platform-as-a-Service (PaaS) technology, fast message queue and schema-free structure-less flexible cloud-ready data-storage solutions.

This solution will relieve end-users from vendor-lock-ins and the rigid costly requirements of maintaining software and hardware infrastructure.

1. Scope of Work (to be done by the student independently)

Data Model:

Initially the Goal is to work with a specific data model like Indian Stock Market and Mutual Funds and show the streams of events (purchase of shares) instantaneously on the UI. Data will be stored and analyzed in Data Warehouse system as soon as they arrive at the Application. Then we need to analyze the data of each share when user drills down a particular stock symbol. ‘Stocks’ are just examples. We envision that the Data Model should be heterogeneous and very dynamic in nature; that means one Entity can have different set of attributes compared to the other.

Data Analysis:

Data Analysis refers to computing SUM, MIN, MAX. AVG, MEDIAN and other aggregate functions. Adopting a flexible cloud architecture allows one to enhance the solution gradually as the data volume grows and implement sophisticated Map Reduce technology.

Data Warehouse:

Data warehousing involves extraction, transformation and loading of incoming data.

We shall use MongoDB as the one-stop-shop for ‘schema-less dynamic object storage’, ‘fast data-analysis provider’ and cloud-ready scalable data-warehouse.

Cloud Enabled Application Platform

In order to cope up with large volume of data and multiple tenants we need a scalable, secure, fail-safe infrastructure which is synonymous with Platform-as-a-Service. We shall pick up a PaaS provider which is the most developer-friendly and innovative solution allowing users to choose from private, public, hybrid and hosted IaaS providers. We want to leverage its powerful capabilities like simple application deployment, automatic resource allocation, automatic service provisioning (an impressive array of databases , message queues, application frameworks and languages.

Future Scope:

Time is ripe for adopting aPaaS (Application Platform as a Service) which is best manifested in the stable and mature platform provided by VMWare-STS-CloudFoundry. Co-existence of several cutting-edge technologies on the same Cloud Platform will allow us easily develop, deploy, share and monitor ‘Large-scale Analytics as a Service’ solution for many different vendors, without worrying about the issues of Security, Multi-Tenancy, Software and Hardware. CEAP not only provides seamless integration between multiple services but also offers high performance, low-latency, fail-over and scalability. Successful implementation of Data warehousing on Cloud will also open up the possibilities of offering services like ‘Data Visualization and Analytics for Big Data’

1. Plan of Work (Work to be done during the semester)

Define the Data Model to start with.

Plan the source of data and design the strategy for Extract, Transform, Load and Analysis of incoming data.

Connect with Cloud Enabled Application Platform (in this case CloudFoundry)

Setup the application development environment and Cloud clients.

Publish incoming events through fast message queue and analyze data using fast

parallel, structure-free dynamic data-warehouse.

Deploy apps to Cloud using client scripts.

Demonstrate the analytics functionalities by computing aggregates (Max, Min, Sum,

Avg) of stock price/volume and displaying the summary report on demand.

1. Literature References

<http://www.infoq.com/presentations/Cloud-Foundry-Design-and-Architecture>

<http://www.mongodb.org/display/DOCS/VMware+CloudFoundry>

<http://en.wikipedia.org/wiki/Cloud_computing>

<http://en.wikipedia.org/wiki/Data_warehouse>

<http://en.wikipedia.org/wiki/Analytics>

1. Particulars of the Supervisor and Examiner
2. Remarks of the Supervisor

2.    Objectives

As a programmer it’s amazing and highly motivating to experience the recent path-breaking innovations in the fields of Cloud Computing which is being adopted by all small, medium, large-scale Enterprises. **Cloud Enabled Integration Platform** (CEIP) and **Cloud Enabled Application Platform** (CEAP) are emerging as the best cloud offspring.

CEIP **integrates** various services (both SaaS and on-premise) on cloud. CEAP offers **rapid application development** on cloud.

In both the cases users do not need to worry about software and hardware infrastructure as all CEIP and CEAP vendors already depend on (homegrown or third-party) stack of cloud technologies (

(Infrastructure as a Service) IaaS, (Platform as a Service) PaaS and (Software as a Service) SaaS [in some cases] ).

This emerging software model opens up the possibilities of implementing many different interesting services and solutions by leveraging Cloud-ready Scalable technologies in all the domains of Database, Messaging, Application Server, Event Processor etc.

One such promising Cloud-enabled Software solution is ‘Data Warehouse Service on Cloud’. This is an open ground with full of possibilities and innovations.

Our objective is to build a Cloud-based Data Warehouse solution for analyzing heterogeneous set of data for diverse domains by leveraging the Platform-as-a-Service (PaaS) technology, fast message queue and schema-free flexible cloud-ready data-storage solutions.

This solution will relieve end-users from vendor-lock-ins and the rigid costly requirements of maintaining software and hardware infrastructure.

3.    Scope of Work (to be done by the student independently)

**Data Model:**

Initially the Goal is to work with a specific data model like Indian Stock Market and Mutual Funds and show the streams of events (purchase of shares) instantaneously on the UI. Data will be stored and analyzed in Data Warehouse system as soon as they arrive at the Application. Then we need to analyze the data of each share when user drills down a particular stock symbol. ‘Stocks’ are just examples. We envision that the Data Model should be heterogeneous and very dynamic in nature; that means one Entity can have different set of attributes compared to the other.

**Data Visualization :**

Web application UI for showing the data being analyzed and the summary reports of data warehouse.

**Data Analysis:**

Data Analysis refers to computing SUM, MIN, MAX. AVG, MEDIAN and other aggregate functions. Adopting a flexible cloud architecture allows one to enhance the solution gradually as the data volume grows and implement sophisticated Map Reduce technology.

 Statistical techniques for generation a robust, consistent data model   
Interactive data exploration/visualization and discovery

**Data Warehouse:**

Data warehouse design (conceptual, logical and physical)

Data warehousing involves Extraction, Cleansing, Transformation and Loading (ETL) of incoming data.

We shall use an advanced database as the one-stop-shop for ‘schema-less dynamic object storage’, ‘fast data-analysis provider’ and cloud-ready scalable data-warehouse.

**Cloud Enabled Application Platform:**

In order to cope up with large volume of data and multiple tenants we need a scalable, secure, fail-safe infrastructure which is synonymous with Platform-as-a-Service. We shall pick up a PaaS provider which is the most developer-friendly and innovative solution allowing users to choose from private, public, hybrid and hosted IaaS  providers. We want to leverage its powerful capabilities like simple application deployment, automatic resource allocation, automatic service provisioning (an impressive array of databases , message queues, application frameworks and languages.

**Future Scope:**

Time is ripe for adopting aPaaS (Application Platform as a Service) which is best manifested in the stable and mature platform provided by VMWare-STS-CloudFoundry.

Co-existence of several cutting-edge technologies on the same Cloud Platform will allow us easily develop, deploy, share and monitor ‘Large-scale Analytics as a Service’ solution for many different vendors, without worrying about the issues of Security, Multi-Tenancy, Software and Hardware.

CEAP not only provides seamless integration between multiple services but also offers high performance, low-latency, fail-over and scalability. Successful implementation of Data warehousing on Cloud will also open up the possibilities of offering services like ‘Data Visualization and Analytics for Big Data’

**Data warehousing consistency and quality   
Data warehouse maintenance and evolution   
Performance optimization and tuning**

**Massive data analytics   
Scalability and parallelization for cloud intelligence: map-reduce and beyond   
Analytics for unstructured, semi-structured, and structured data   
Analytics for data streams  
Real-time analytics   
Privacy and security in cloud intelligence   
Reliability and fault tolerance in cloud intelligence**

4.    Plan of Work (Work to be done during the semester)

Define the Data Model to start with.

Plan the source of data and design the strategy for Extract, Transform, Load and Analysis of incoming data.

     Connect with Cloud Enabled Application Platform  (in this case CloudFoundry)

     Setup the application development environment and Cloud clients.

     Publish incoming events through fast message queue and analyze data using fast  parallel, structure-free dynamic data-warehouse.

     Deploy apps to Cloud using client scripts.

     Demonstrate the analytic functionalists by computing aggregates (Max, Min, Sum,

     Avg) of stock price/volume and displaying the summary report on demand.