CSYE 7250 Big Data Architecture & Governance

Final, Term Project

Fall18

[Data Siloes Integration]

Submitted by,

Andrews Tito

001837549

**Table of Contents**

[**1.** **Introduction** 3](#_Toc532456053)

[**2.** **Functional requirements** 3](#_Toc532456054)

[**3.** **Non-functional requirements** 4](#_Toc532456055)

[**4.** **Vision diagram** 5](#_Toc532456056)

[**5.** **Strategy & Architecture** 5](#_Toc532456057)

[**6.** **Project Plan** 7](#_Toc532456058)

[**7.** **Other Factors** 8](#_Toc532456059)

[**7.1** **Issues & Risk** 8](#_Toc532456060)

[**7.2** **Data collection** 8](#_Toc532456061)

[**7.3** **Security** 9](#_Toc532456062)

[**7.4** **Scalability** 9](#_Toc532456063)

[**7.5** **Management** 9](#_Toc532456064)

# **Introduction**

The data integration project enables to provide a streamlined data dashboard for medical providers and research team, to provide access to integrated datasets that will support evidence-driven clinical decision making. It helps in provide new insights for adapting standardized personalized healthcare protocols to individuals based on their genetic makeup and likely responses to specific therapies. Integration of genomic data into the clinical workflow and collection and analysis of Molecular Information for diagnosing and treating diseases, by going beyond traditional diagnostics.

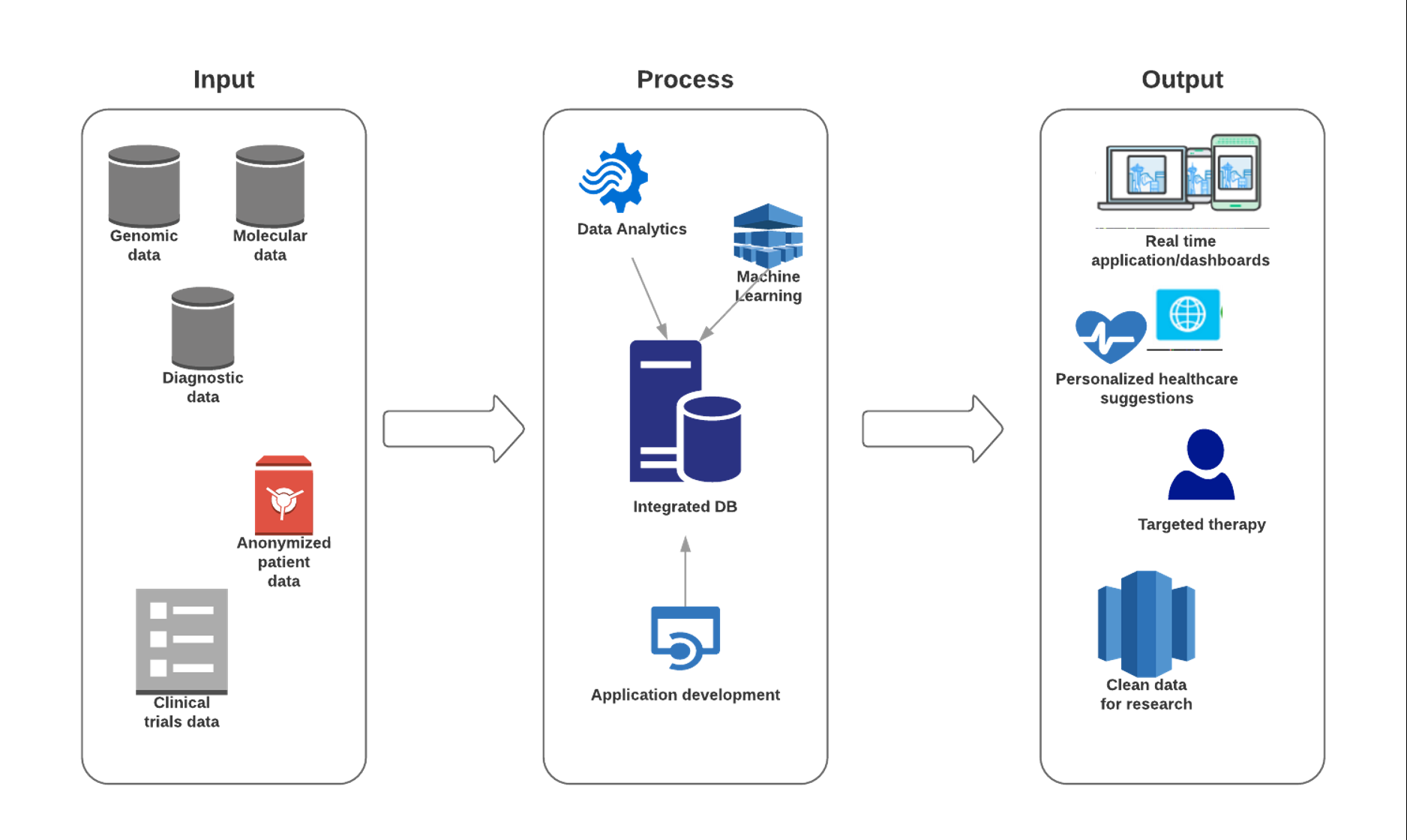
# **Functional requirements**

* **Interoperability**: The key consideration for any big data solution will be whether the solution would work with multiple platforms or a single platform. Considering a baseline functionality, the requirement would be to build a closed analytics application, that is built for one infrastructure platform. However, for the project, considering the business opportunity of a licensing system, as a best practice functionality, an analytics solution that would interface with any of the major platforms is recommended considering a future proof implementation.
* **Asset visualization:** Processed and relevant data will be available for researchers and other beneficiaries. These data need to be presented in a visual form that is easily understandable across all beneficiaries. Considering a baseline functionality, raw statistical and integrated information requires further analysis from big data scientists. However, for the project, as a best practice functionality, high quality Business Intelligence tool with company wide visibility is a must.
* **Batch processing:** As the data sets are large, often a big data solution must process data files using running batch jobs to filter, aggregate, and prepare the data for analysis. Usually these jobs involve reading source files, processing them, and writing the output to new files. As for the project, as a best practice functionality, it is important to process large volumes of data to get more detailed insights as well as to get fast analytics results.
* **Integrated database:** Data for batch processing operations is typically stored in a distributed file store that can hold high volumes of large files in various formats. This kind of store is often called a data lake. For the project, as a best practice functionality, the integrated system would allow storage of a virtually unlimited amount of data, cope both with high rates of random write and read access, flexibly and efficiently deal with a range of different data models, support both structured and unstructured data, and only work on encrypted data.
* **Robust Framework & Orchestration:** Any kind ofBig Data frameworks is designed to cope with the problems of Variety, Volume and Velocity. Key features allow for iterative processing, storage and data ingestion. For the project, as a best practice functionality, the framework should allow repeated data processing operations, encapsulated in workflows, that transform source data, move data between multiple sources and load the processed data into an analytical data store, or push the results straight to a report or dashboard generating tool.

# **Non-functional requirements**

* **Performance** – This requirement ensures optimal performance for the Big data solution, by considering the following:
  + Performance consideration for Data Acquisition
  + Performance consideration for Storage
  + Performance consideration for Data Processing
* **Reliability –** The solution should ensure that the system should continue to work correctly (performing the correct function at the desired level of performance) even in the face of adversity (hardware or software faults, and even human error).
* **Maintainability -** Over time, many different people will work on the system and they should all be able to work on it productively by ensuring that:
  + Monitoring the health of the system periodically
  + Keeping software and platforms up to date, including security patches
  + Establishing good practices and tools for deployment, and configuration management
* **Cloud Support & Elasticity –** The requirement ensures that the solution should be capable to move the architecture into the cloud to take advantage of the many benefits associated with its economies of scale. Elasticity is a cloud property which allows a system to scale up and down according to demand.
* **Flow Control –** The requirement ensure that the solution is capable to handle scenarios where the data source is emitting records faster than system can consume by providing strategies like dropping records and sampling.

# **Vision diagram**



# **Strategy & Architecture**

1. Initial step involves **collection of initial data** and exploratory analysis of the data properties such as separating data into subsets in order to form and evaluate hypotheses. Here the data presented are a mix of structured, and semi-structured data and special data handling procedures are carried out due to the sensitivity of the data being used. Once the all the data points are finalized, it is passed on to the Data Analysts and other personnel for further processing of the data.
2. For **Data Analytics**, the project will be using Tableau as a Business Intelligence tool for visual analysis of the data. Its main advantage is its ability to connect to multiple data sources like file-based data sources, relational databases, OLAP data sources and cloud-based data sources which is a key factor in handling multiple data points initially in the project. Tableau also helps the project to achieve interoperability and asset visualization by being accessible to all levels of organizations and individuals based on authorization. It also provides enterprise-class business analytics platform that can scale up to hundreds of thousands of users by mobile and browser-based analytics.
3. For the purpose of **Integrated Database**, HBase, a non-relational data store for Hadoop is used. It features linear and modular scalability, strictly consistent reads and writes, and automatic failover support. It has easy java API for client and provides data replication across clusters. It manages to fulfill the projects need to cope with high rates of random write and read access. Also, HBase is an excellent database for structured and semi-structured data, which is the main type of data the project will be handling.
4. **Machine learning** in the project can be depicted as follows:

Genetics and Genomics - The research in genetics and genomics enables an advanced level of treatment personalization. The goal is to understand the impact of the DNA on our health and find individual biological connections between genetics, diseases, and drug response. MapReduce allows reading genetic sequences mapping and shortens the time for efficient data processing. SQL contributes to retrieving genomic data, file manipulations, and computation.

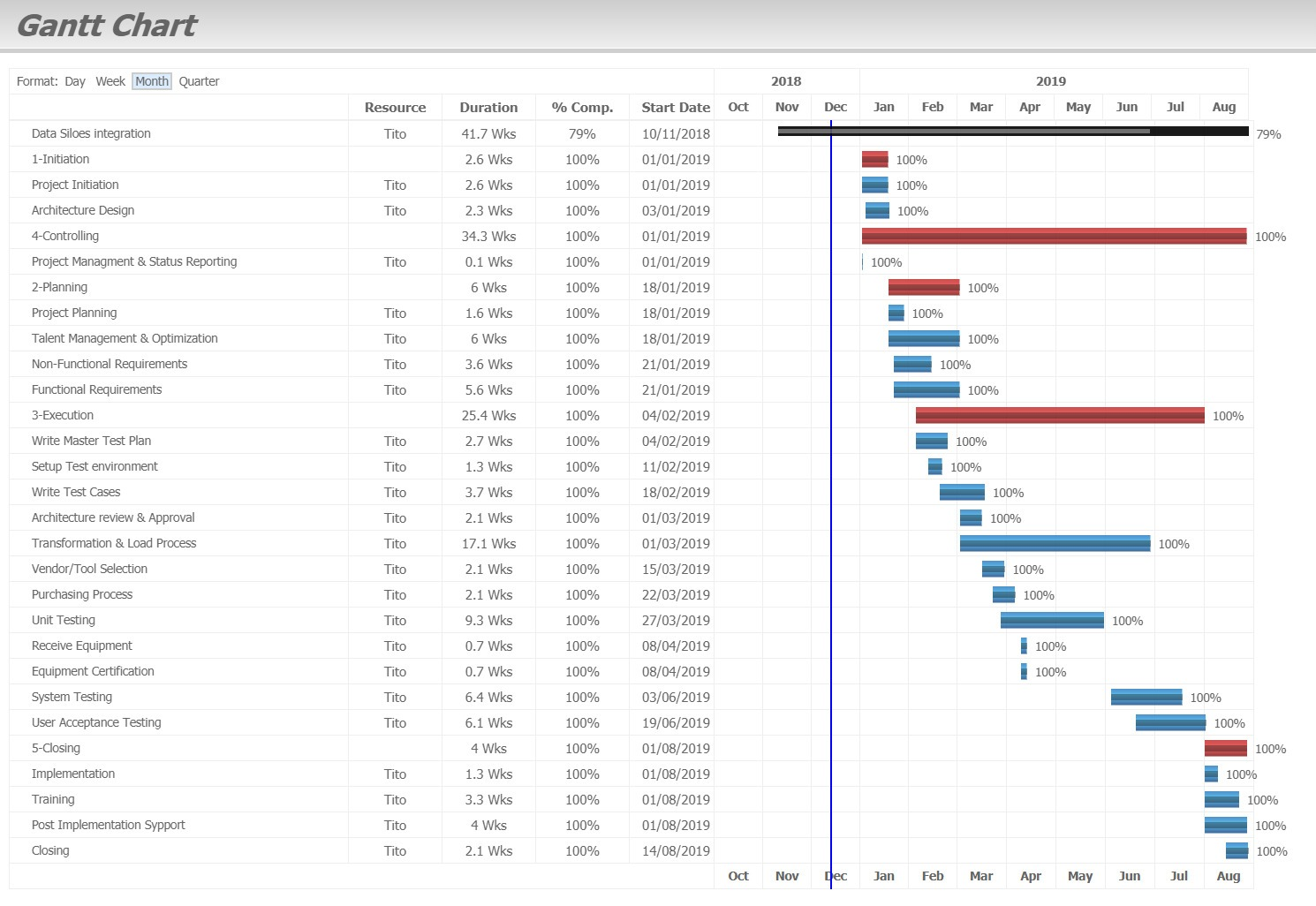
Drug discovery - The drug discovery process is highly complicated and involves many disciplines. On average, it takes twelve years to get a drug officially submitted. The machine learning algorithms simplify and shortens this process. These algorithms can forecast how the compound will act in the body using advanced mathematical modeling and simulations instead of the “lab experiments”. The computational drug discovery also improves the collection and application of different types of historical data during the drug development process.

Predictive diagnostic accuracy - The predictive analytics methods learn from historical data and make accurate predictions about the outcomes. They process the patient data, make sense of clinical notes, find the correlations, associations of symptoms, habits, diseases, and then make predictions. The impacts of certain biomedical factors such as genome structure or clinical variables are taken into the account to predict the evolution of certain diseases.

1. The **framework** for the project mainly will be Apache Hadoop. Hadoop is open source and Java-based programming framework. Data transfer between web-based clients are protected by SSL thereby providing an additional layer of security to the project. 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. 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. Hadoop is also a highly scalable storage platform, because it can store and distribute very large data sets across hundreds of inexpensive servers that operate in parallel.
2. The **network infrastructure** involved in the data integration project should have the following key abilities:

* ability to handlle unpredictable content with no apparent schema or structure
* must address a traffic flow shift from the server-to-client pattern to a heavier server-to-server traffic flow across the datacenter
* to be able to withstand constant flow of workload-intensive transactions

# **Project Plan**



*(full details available in Velero PPM)*

|  |  |  |
| --- | --- | --- |
| Category | Start date | End date |
| 1 - Project Initiation | 1-Jan-19 | 18-Jan-19 |
| 2 - Planning | 18-Jan-19 | 28-Feb-19 |
| 3 - Execution | 4-Feb-19 | 31-Jul-19 |
| 4 - Controlling | 1-Jan-19 | 28-Aug-19 |
| 5 - Closing | 1-Aug-19 | 28-Aug-19 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **PLC Category** | **Type** | **Milestone/Task Description** | **%Complete** | **Start Date** | **End Date** | **Status** |
| 1-Initiation | Analysis | Project Initiation | 100.00% | 1/1/2019 | 1/18/2019 | Complete |
| 1-Initiation | Analysis | Architecture Design | 100.00% | 1/3/2019 | 1/18/2019 | Complete |
| 2-Planning | Analysis | Talent Management & Optimization | 100.00% | 1/18/2019 | 2/28/2019 | Complete |
| 2-Planning | Analysis | Non-Functional Requirements | 100.00% | 1/21/2019 | 2/14/2019 | Complete |
| 2-Planning | Analysis | Functional Requirements | 100.00% | 1/21/2019 | 2/28/2019 | Complete |
| 2-Planning | Analysis | Project Planning | 100.00% | 1/18/2019 | 1/28/2019 | Complete |
| 3-Execution | Analysis | Architecture review & Approval | 100.00% | 3/1/2019 | 3/15/2019 | Complete |
| 3-Execution | Analysis | Vendor/Tool Selection | 100.00% | 3/15/2019 | 3/29/2019 | Complete |
| 3-Execution | Milestone | Purchasing Process | 100.00% | 3/22/2019 | 4/5/2019 | Complete |
| 3-Execution | Milestone | Receive Equipment | 100.00% | 4/8/2019 | 4/12/2019 | Complete |
| 3-Execution | Milestone | Equipment Certification | 100.00% | 4/8/2019 | 4/12/2019 | Complete |
| 3-Execution | QA | Write Master Test Plan | 100.00% | 2/4/2019 | 2/22/2019 | Complete |
| 3-Execution | QA | Write Test Cases | 100.00% | 2/18/2019 | 3/15/2019 | Complete |
| 3-Execution | QA | Setup Test environment | 100.00% | 2/11/2019 | 2/19/2019 | Complete |
| 3-Execution | Development | Transformation & Load Process | 100.00% | 3/1/2019 | 6/28/2019 | Complete |
| 3-Execution | QA | Unit Testing | 100.00% | 3/27/2019 | 5/30/2019 | Complete |
| 3-Execution | QA | System Testing | 100.00% | 6/3/2019 | 7/17/2019 | Complete |
| 3-Execution | QA | User Acceptance Testing | 100.00% | 6/19/2019 | 7/31/2019 | Complete |
| 4-Controlling | Development | Project Management & Status Reporting | 100.00% | 1/1/2019 | 8/28/2019 | Complete |
| 5-Closing | Next Steps | Training | 100.00% | 8/1/2019 | 8/23/2019 | Complete |
| 5-Closing | Systems | Implementation | 100.00% | 8/1/2019 | 8/9/2019 | Complete |
| 5-Closing | Systems | Post Implementation Support | 100.00% | 8/1/2019 | 8/28/2019 | Complete |
| 5-Closing | Milestone | Closing | 100.00% | 8/14/2019 | 8/28/2019 | Complete |

# **Other Factors**

## **Issues & Risk**

|  |  |
| --- | --- |
| Issues | Risk |
| 1. Incompatible Legacy data | 1. Hard drive failure |
| 2. Inconsistent data types | 2. Network load imbalance |
| 3. Project plan execution delay | 3. Late UAT testing |
| 4. Employee vacation time | 4. Database connector failure |
| 5. Late supply of equipment’s | 5. Policy changes on usage of patient data |

## **Data collection**

The main sources of data for this project is Genomic data, Molecular data, Diagnostic data, Clinical trials data and other anonymized patient data. All these data points exist as siloes which will be integrated in the project. Although even after integration is done, the data siloes still can be used feed updated data, as and when it becomes available. Through integration, the intention to achieve data to support evidence-driven clinical decision making based on their genetic makeup and likely responses to specific therapies. Also anonymized patient data is grouped to identify new potential targets for therapy. When the data package hits the data warehouse, it is stored in different formats, the best format is the plain text format as it is very flexible. For bigger and more structured companies SQL or any other structured format could work.

## **Security**

Following steps are taken in order to manage security between all architecture layers:

* Encrypt all outgoing data for protection
* Access permissions for system data may only be changed by the system’s data administrator
* Secure all in-network connected devices with device management policies
* Introduce identity and access management (IAM)
* Continuous monitoring of data usage
* Regular system backup to be performed and the backup copies needs to be stored in a secure location
* Apply fine-grained access controls for people accessing patient data
* Apply individual data ownership and control to tackle day to day issues
* Eradicate departmental data silos to avoid application centric approach

## **Scalability**

**Scalability** refers to the capability of a system to handle a growing amount of work when processing power is expanded to accommodate growth. A system is said to be scalable if it can increase its workload and throughput when additional resources are added. Developing a comprehensive scalable data platform is key to continuing the projects development. Elasticity is a property which allows a system to scale up and out according to demand. **Scaling up or Vertical scaling** is process in which you add new resources like memory or CPU to improve system performance. **Scaling out or Horizontal scaling** is the process of adding more hardware to a system.

When increase in number of data points and increase in storage – **Scale out**

When increase in number of transactions – **Scale up**

## **Management**

Manageability is brought about by using **Apache ZooKeeper**. ZooKeeper is a distributed application providing services for managing a distributed application. In the project, ZooKeeper helps to coordinate and manage shared data with robust synchronization techniques. Some common services provided by the ZooKeeper are Naming service, Configuration management, Cluster management, Locking and synchronization service and reliable data registry. ZooKeeper uses a custom messaging protocol, which ensures that the local replicas never diverge. It also provides high performance by its reads outnumber writes, which is a typical showcase for a coordination service. It is also able to sustain a high throughput even when a server node which performs automatic recovery fails. One main highlight of this is that, no transaction is left partial, rather, it succeeds completely or fails completely which avoids any data corruption in the process.