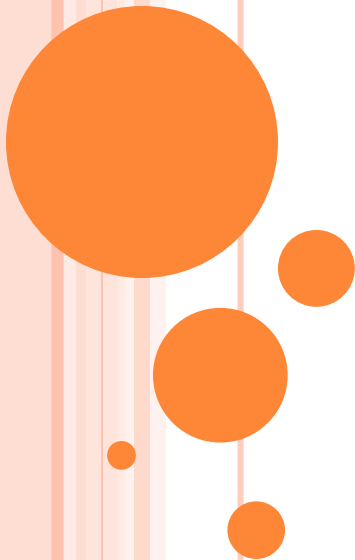


Real-Time Analytics on BigData

(Final Review Report)

Sakthivel M
2012HZ78469
Cognizant



Objective

- ▶ The objective of this dissertation work is to build a Real-Time analytics system that stores and process large volume of data and also capable of answering to the queries on an ad-hoc basis.
- ▶ The aim is to improve the batch processing capacity of the existing BigData systems like Hive, with real-time or near real-time querying capabilities.



Scope

- ▶ The scope of this dissertation is to build a real-time analytics on HL7 data stored in big data systems. HL7 Messages are used to transfer electronic data between disparate healthcare systems.
- ▶ Streaming of HL7 messages will be stored in Hadoop platform. It can then be later processed with Hive to make use of its batch processing capability.



Scope(Contd.)

- ▶ The HL7 message will also be processed with Apache Spark, the system capable of doing in-memory computations to get low latency query result.
- ▶ HL7 supports wide range of information exchange/transfer of clinical and administrative data between Hospital information systems(HIS). But, this system at this point in time is going handle limited set of activities like admission, discharge, etc..
- ▶ The comparison of both the system's capability will be visually represented with graph.



Benefits of the dissertation

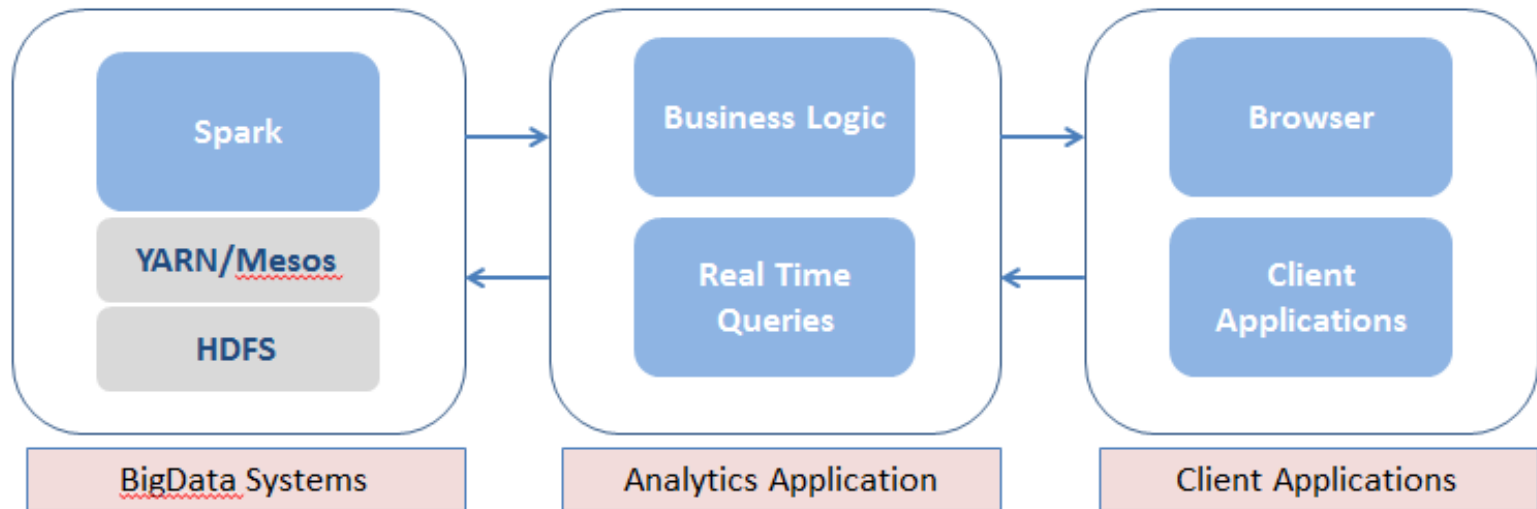
- ▶ Most of the BigData processing systems basically supports batch processing by default. By adding the low latency querying capability will cater the demands of client business need such as real-time data access.
- ▶ We can leverage this solution with the ongoing and emerging technology of BigData solution development
- ▶ As Cognizant starts working on innovative ideas and emerging technologies like Social, Mobile, Analytics and Cloud (SMAC), this can be very well used alongside with any of the technologies.



Project plan

Stage	Purpose	Activities	Status	Deliverable
Stage 1	<ul style="list-style-type: none"> Define Scope of Dissertation Define Requirements 	Elicitation of Business Process	100%	Preliminary Report
		Define software requirements (SRS) <ul style="list-style-type: none"> Functional requirements in use case format Non-functional requirements 	100%	
Stage 2	<ul style="list-style-type: none"> Propose Solution Detailed design for the proposed Solution 	Proposed Solution <ul style="list-style-type: none"> High Level design Technical Requirement – Software & Hardware for development and production implementation 	100%	Mid Review report
		Detailed design <ul style="list-style-type: none"> Control flow Data flow Development & Testing considerations 	70%	
Stage 3	QA and Documentation	<ul style="list-style-type: none"> Development & Unit testing QA testing & Assessment 	50%	Final review Report
		Define future extensibility opportunities	10%	

Solution Architecture



Solution Architecture

Apache Hadoop:

- Apache Hadoop is an open-source software framework for storage and large-scale processing of data-sets on clusters of commodity hardware is composed of the following modules:
- Hadoop Distributed File System (HDFS) – a distributed file-system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster.
- Hadoop YARN – a resource-management platform responsible for managing compute resources in clusters and using them for scheduling of users' applications.
- Hadoop MapReduce – a programming model for large scale data processing.



Solution Architecture

Apache Hive:

- Apache Hive is a data warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis.
- It provides an SQL-like language called HiveQL while maintaining full support for map/reduce. To accelerate queries, it provides indexes, including bitmap indexes.
- SQL-like queries (Hive QL), which are implicitly converted into map-reduce jobs.



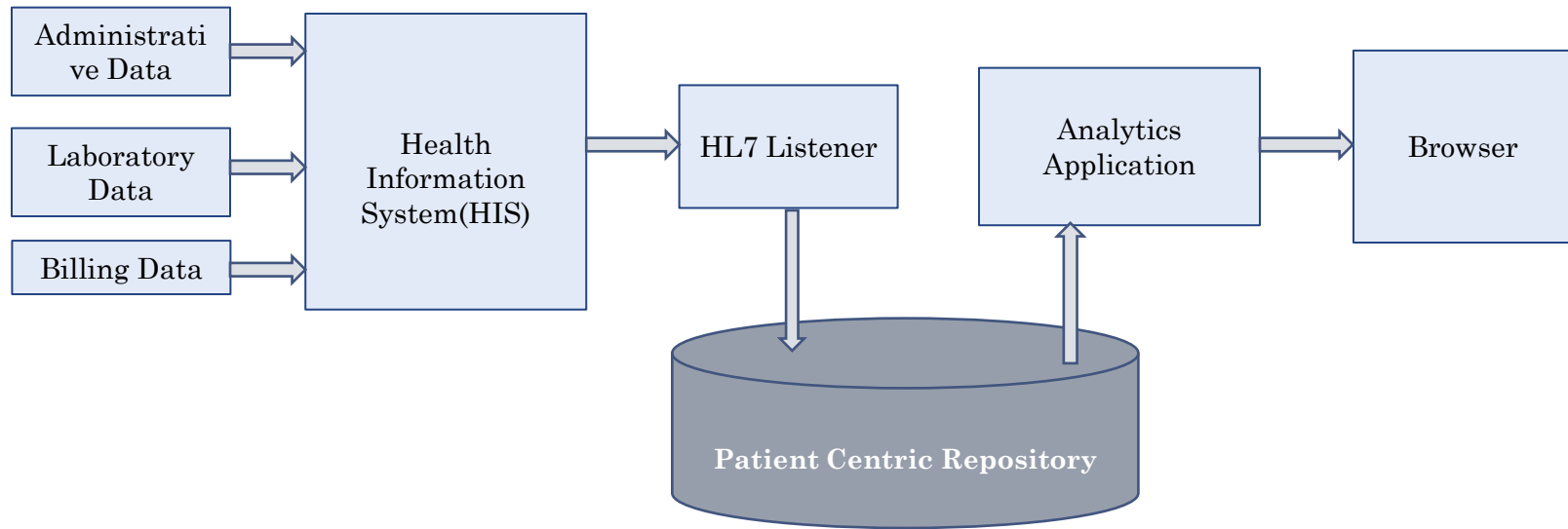
Solution Architecture

Apache Spark:

- Apache Spark is an open-source data analytics cluster computing framework.
- Spark fits into the Hadoop open-source community, building on top of the Hadoop Distributed File System (HDFS).
- However, Spark is not tied to the two-stage MapReduce paradigm, and promises performance up to 100 times faster than Hadoop MapReduce for certain applications.
- Spark provides primitives for in-memory cluster computing that allows user programs to load data into a cluster's memory and query it repeatedly



Data Flow Diagram



Tools and Techniques used

- **BigData Platforms**
 - Hadoop, HBase, Spark
- **Platform:**
 - Java, Scala
- **Framework used:**
 - Play
- **Webserver:**
 - JBoss Netty
- **Testing:**
 - JUnit
- **Development environment(IDE):**
 - Eclipse



Screenshots - Signup

Sign Up to Analytics App


Sign Up

[Sign In](#)



Screenshots - Signin

Sign in to continue to Analytics App



sakthi@gmail.com|

.....

Sign in

☐ Remember me [Need help?](#)

[Create an account](#)



Screenshots – Home Page

Real-time

Real-time refers to a level of computer responsiveness that a user senses as immediate or nearly immediate, or that enables a computer to keep up with some external process (for example, to present visualizations of Web site activity as it constantly changes). It consists of dynamic analysis and reporting, based on data entered into a system less than one minute before the actual time of use. Real-time analytics is also known as real-time data analytics, real-time data integration, and real-time intelligence.

In this context, real-time means a range from milliseconds to a few seconds after the business event has occurred.



Monitor as it happens



Analytical Dashboard

The visualization component then reads the data from the structured data file (JSON/XML), and draws a chart, gauge, or other visualization in the reporting interface. The frequency at which



Screenshots – Patient Search

Patient Search

Comparison

Enter patient id



All

Admission

Discharge

Submit

S.No	Date	Type	PatientId	Name	DoB	Sex	Address	Married	SSN No
1	02-10-2014	Admission	P1000	Maria Anders	14-09-1985	F	1200 N ELM STREET, GREENSBORO, NC	Y	078-85-7120
2	02-10-2014	Admission	P1001	Christina Berglund	14-09-1998	F	Maddison Square, NY	N	078-05-1520
3	02-10-2014	Discharge	P1002	Francisco Chang	14-09-1985	M	123 Times Square, Whites Road, CA	Y	078-15-1120
4	02-10-2014	Admission	P1003	Roland Mendel	14-09-1995	M	1288 Anderson Garden, GREENSBORO, DC	N	090-05-1620
5	02-10-2014	Discharge	P1004	Helen Bennett	14-09-1985	M	1200 N ELM STREET, GREENSBORO, NC	Y	079-25-1720



Screenshots Patient Search-Id

Patient Search

Comparison

P1000



All

Admission

Discharge

Submit

S.No	Date	Type	PatientId	Name	DoB	Sex	Address	Married	SSN No
1	13-11-2014	Admission	P1000	Maria Anders	14-09-1985	F	1200 N ELM STREET, GREENSBORO, NC	Y	078-85-7120
1	12-10-2014	Discharge	P1000	Maria Anders	14-09-1985	F	1200 N ELM STREET, GREENSBORO, NC	Y	078-85-7120
1	05-10-2014	Admission	P1000	Maria Anders	14-09-1985	F	1200 N ELM STREET, GREENSBORO, NC	Y	078-85-7120

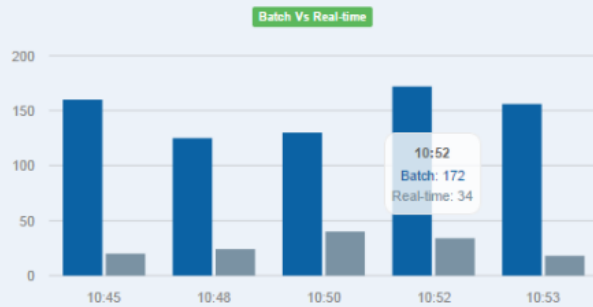


Screenshots - Comparision

Patient Search

Comparison

Latest Comparison History(Last 5 Entries)



Thank You

