Leading a Healthcare Company to the Big Data Promised Land:

A Case Study of Hadoop in Healthcare

Mohammad Quraishi (IT Senior Principal - Cigna)
atif71@gmail.com
About me

• BS in Computer Science and Engineering from University of Connecticut
• In the Healthcare Industry for over 19 years
  • Programmer most of my career - Architect, Designer
  • Worked in the SOA space for a number of years
  • Lead engineer in the mobile application space
  • Now Lead engineer in the Big Data Analytics Space - Hadoop

In my spare time
• Love to travel with the family
• Video games, music, movies
• Community relations work
• Fan of College basketball
Breakdown of the Hadoop Journey

1. Making the case
   Vision
   Architecture

2. The blowback
   What we accomplished

3. Roadmap to the future
   Lessons Learned
   Questions?
The Elephant in the room

Image Credit: Guian Bolisay/Flickr
What’s the problem?

We already have a mature data analysis infrastructure
And it looks something like this...

What we already do

- We have independent data marts
- We have the Hub-and-spoke architecture, the centralized warehouse
What is the vision?

The ability to perform
• Descriptive, Predictive and Prescriptive Analytics

Remove the traditional IT barriers separating the business users from insights
Benefits of Big Data

• Hadoop has the lowest cost per TB ratio of any data technology available
• Getting started with Hadoop is fairly inexpensive
  • “Entry-level” clusters relatively inexpensive
  • Grow in small steps
Benefits of Big Data

You don’t have to throw away data anymore!
Vision - Reference Architecture

1. Hadoop Cluster running HDFS and MapReduce
   Includes Management, Monitoring and Security

2. Edge Node For Hadoop Client
   - SQOOP/Flume Chronos
   - Python
   - Hive/Impala (SQL)
   - Scalding (Scala)
   - Cascading (Java)

3. Realtime Data Store or event processing
   - Logs Web
   - IVR
   - Portal Mobile
   - Storing weblogs
   - Log files to HDFS

4. External Hadoop Output
   - Teradata
   - RDBMS
   - Back up or copy data from HDFS to a redundant cluster for quick recovery
   * For future implementation TBD

5. CED/Claims
   - Clinical Data
   - RDBMS

6. Data Science Tools
   - Tableau
   - Spotfire
   - Platfora
   - Cognos
   - Microstrategy

7. Analysis/Modeling Tools
   - SAS
   - Pentaho
   - R

8. Teradata
   - filestack
   - Analysis
   - SQOOP/Flume
   - Chronos
   - Hadoop Cluster #2
   - Hadoop Cluster #3
   - Jobs
   - MapReduce Distributed Programming Framework
   - Streaming and append to Hadoop output - Realtime events

9. Realtime Feed
   - Flume
   - FLume
   - SQOOP
   - SQOOP/Flume
   - Chronos
   - Hadoop Cluster #2
   - Hadoop Cluster #3
   - RDBMS
   - *Use Spark Streaming and append to Hadoop output - Realtime events

10. Web Analytics
    - Event detection(Storm)

11. Hadoop Cluster running HDFS and MapReduce
    Includes Management, Monitoring and Security

12. Web Analytics
    - Event detection(Storm)

13. Hadoop Cluster running HDFS and MapReduce
    Includes Management, Monitoring and Security

14. Web Analytics
    - Event detection(Storm)
The Initial Evaluation

• Vendor Evaluation: Which relationship best fits our needs without lock-in?

• Selection of use cases for demonstration

• Visualization of those use cases
Use Case 1
Success!

• Ready to tackle tougher more complicated problems

• Went out looking for more use cases
Ran into misconceptions

“Let’s use Hadoop as ETL!”

“Help us move data.”

“Can we back up data for archiving?”
... & Challenges
But Why?

• Overuse of the words “Big” & “Data”

• There was an overlap with other tools and platforms

• Hadoop looked like a swiss army knife

• Will it take over the world and replace other platforms?
Broader impact - Business Benefits

• Building a Customer Persona
• Service Ops efficiency
• Being Customer Centric
• Product Efficiency
• Brand Impact
Broader impact - IT Benefits

• Predictive threat modeling
• Data Archival
• Network Efficiency
Hadoop and Big Data

• Big Data = Hadoop + Relational + other suitable task related technologies

• Hadoop is complementary
Hadoop is Complementary

• Hadoop excels at processing and analyzing large volumes of distributed, unstructured, structured and semi-structured data in batch or near real-time fashion for analysis.

• NoSQL databases are adept at storing and serving up multi-structured data in near-real time for web-based applications.

• Massively parallel OLAP databases are best at providing analysis of large volumes of mainly structured data - Teradata.

• SAS/R - Modeling and Business Intelligence.

• Tableau - Visualization.
Embrace the Most Important Change: *Culture*

*Democratize your data and reap the benefits!*
## Why is Hadoop Complementary?

<table>
<thead>
<tr>
<th></th>
<th>Hadoop</th>
<th>Relational</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Analysis type</strong></td>
<td>Exploratory analysis to uncover value in data</td>
<td>Operational analysis of what was uncovered</td>
</tr>
<tr>
<td><strong>Data granularity</strong></td>
<td>Store High Volumes of Highly Granular data – lowest level; disk is cheap</td>
<td>Store transformed, aggregated data – conserve processing and storage costs</td>
</tr>
<tr>
<td><strong>Time frame</strong></td>
<td>Volumes and Varieties of data that is analyzed is streamed directly into Hadoop</td>
<td>Long term trending analysis from data that is provided by utilizing Hadoop</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th><strong>Hadoop</strong></th>
<th><strong>Teradata</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Maturity</strong></td>
<td>Rapid evolution. Documentation and tooling are rough around the edges.</td>
<td>Stable, mature system.</td>
</tr>
<tr>
<td><strong>Cost</strong></td>
<td>Lowest $/GB available.</td>
<td>~10-100x the cost of Hadoop.</td>
</tr>
<tr>
<td><strong>Data Model</strong></td>
<td>Full spectrum from relational to unstructured, i.e., suitable for queries to machine learning problems.</td>
<td>Relational only.</td>
</tr>
</tbody>
</table>
What we accomplished?

• Evangelized Hadoop

• Linked Hadoop to BI Tools

• R on Hadoop

• A fail fast iterative analytics approach
Lambda Architecture as the foundation

The master dataset is the only part of the Lambda Architecture that absolutely must be safeguarded from corruption. So for this reason the fault tolerance and redundancy of HDFS adds tremendous support. The master data set is also referred to as "Raw Data" or "Bronze Data" in our reference architecture and wiki in general. Information created in the serving and speed layers is also referred to as "Silver Data".

Reference Implementation

The implementation of the reference architecture is shown below. Over time Hadoop will interact with various BI tools and other technologies and reference architecture assist in multiple domains that cover, ETL (for data cleansing, data movement), analytics, archival, log parsing, realtime convergence of data, assistance in the BI space and much more. Below is one representation of this interaction.

Credit Nathan Marz - Big Data
What we accomplished?

- ETL - Ingest, Transform and Move patterns
- Logs generated from consumer channels were ingested with Flume
- Standardized on Parquet (Storage) and Snappy (Compression)
- Lifecycle and organization of Data on HDFS
- LUKS - dm-crypt — for data at rest encryption
- Sentry and LDAP for Role Based Access Control
A Custom NLP Framework

1. **Phonology**
   - Speech analysis
   - Speech synthesis
   - Pronunciation model

2. **Morphology**
   - Morphological analysis
   - Morphological realization
   - Morphological rules

3. **Syntax**
   - Parsing
   - Syntactic realization
   - Lexicon and grammar

4. **Semantics**
   - Contextual reasoning
   - Utterance planning
   - Discourse context

5. **Reasoning**
   - Application reasoning and execution
   - Domain knowledge
A Roadmap to the Future

- **Large Scale Data/ Raw**
  - History Archive
  - MDM sources
  - Other Trxn data
  - Structured Data
  - Social (twitter etc)
  - Audio
  - Web, IVR
  - Mobile Logs
  - Live Data
  - Click stream
  - Web Analytics
  - CRM
  - Campaign
  - Marketing
  - Source Data
  - Unstructured
  - Semi-structured

- **Real-time/Batch Analytics**
  - Spark Core
    - Spark Streaming
    - Spark SQL (Hive)
    - MLib
  - Views
    - Batch/Realtime Views
    - Data obfuscation
    - Create views for BI tools

- **Business Processes Applications**
- **Consumer Channels/Apps**
  - Mobile Apps
  - Portals

- **ETL**
- **Low Level RESTful Services**
  - Built with Akka

- **External Repos**
  - RDBMS
  - NoSQL

- **Insights Dataset**
  - ETL Integration

- **Relational Data**
  - From Hadoop or SAS/R
  - Organizational MDM Sources

- **MPP High Performant DB**

- **BI**
  - SAS
  - R
  - Tableau
  - R

- **BI and Predictive Analytics**
  - Visualization/Data Science Tools
A Roadmap to the Future

Data Driven Solutions + FP

“Functional Programming: I came for the concurrency, but I stayed for the Data Science”
Dean Wampler
There’s also Workflow Management with Oozie.
Lessons Learned

- Overuse of the words “Big” & “Data”
- The overlap
- Everyone found a use for Hadoop
- Big Change/Baby Steps
- Agility + Process = Cognitive Dissonance
Healthcare company needs

• Security

• Vendors

• Vendor Partnerships
“Difficult to see. Always in motion is the future...”

Yoda

“Many of the truths that we cling to depend on our point of view.”

Yoda

The Journey of a thousand miles begins with one cluster...
Questions?

Mohammad Quraishi (IT Senior Principal - Cigna)
atif71@gmail.com