Transition from an Oracle DBA to Big Data architect

Saurabh K. Gupta
@saurabhkg
Who am I?

• Database Leader, Data and Analytics at GE
• 10 years of experience in data engineering, architecture, Oracle technologies
• Authored couple of books with Packt Publishing
  • Oracle Advanced PL/SQL Developer Professional Guide
• Twitter @saurabhkg
Why I’m here?

• As a classical Oracle Database administrators, you know how to deal with fat data sets already. Big data is little different as more than its size, what matters is the variety and velocity.

• Businesses are staking a lot to find the data nuggets out of noisy heaps. There is a lot that DBAs can contribute in this shift. Not just the data availability, but DBAs can transform themselves into data architects by stepping out of classical database administration skills.

• This session will focus on skill areas that can help Oracle DBAs to emerge as Big Data DBAs. The talk will cover the overview of big data ecosystem, key Big Data technologies and what DBAs can leverage from their current skill set to focus on big data DBA.
Agenda

• Big Data – making sense out of nonsense
• How to design a Big Data solution?
• Big Data solution spectrum
• Build you Big Data team
Big Data – Making Sense out of Nonsense
Big Data – Making Sense out of Nonsense

- Structured and unstructured data that augments a business on daily basis –
  - Large volumes of data
  - At a relative velocity
  - With relative variety
  - Can reveal nuggets of information

- Information is more important than Data
  - What we do with the data

- New term; not the concept
  - Data gathering, storage, and analysis has been for a while
Big Data – what Industry thinks?

• "Data really powers everything that we do." – Jeff Weiner, LinkedIn.

• "You can have data without information, but you cannot have information without data." - Daniel Keys Moran

• “Data beats emotions.” – Sean Rad, founder of Ad.ly

• “Hiding within those mounds of data is knowledge that could change the life of a patient, or change the world.” – Atul Butte, Stanford

• “Torture the data, and it will confess to anything.” – Ronald Coase, Economics, Nobel Prize Laureate
How much we love Data

• Facebook (http://newsroom.fb.com/company-info/)
  • 1.18 billion daily active users on average for Sep’16

• Twitter (https://about.twitter.com/company)
  • 313M active monthly users; 82% active users on mobile

• Instagram (https://www.instagram.com/press/)
  • 4.2 Billion likes daily; 30+ Billion in ~5 years

• Google searches
  • 57,115 Google searches in 1 second

• Digital universe will grow to 44 zettabytes (Trillion GB)
How much we love Data

• “Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone.”

•- IBM
How to design a Big Data solution
Plan and develop Big Data solution

Planning
- Identify the challenges
- Solution strategy (develop and deploy)
- Predict the data growth
- Plan your Infrastructure

Solutioning
- Identify data sources and consumers
- Data ingestion strategy
- Data acquisition strategy
- Data Analytics
- Data Consumption
- Operational strategy

AIOUG Sangam’16
Big Data solution spectrum

Data Sources
- DWH
- Image
- Audio
- Video
- WWW

Data Acquisition
- Data collection
- Data transmission
- Pre-processing

Data Storage
- Distributed File System
- Transactional Stores
- EDW

Data Consumption
- BI Tools
- Predictive Analytics
- Data Visualization

https://hadoopecosystemtable.github.io/

AIOUG Sangam’16
Data Acquisition and Ingestion

• Design strategy for data collection, data transmission, data pre-processing
• Understand the nature of data
  • Data gen rate, volume, batch or stream
• Data ingestion tools - Sqoop, Flume, Kafka, Storm
• Web crawling tools (Apache Nutch and open-source)
• Oracle GoldenGate for Big Data 12c

AIOUG Sangam’16
Data Acquisition and Ingestion

• Flume - Distributed system for collecting and aggregating log data, and writing it to HDFS. Simple, flexible, and highly available. Tightly integrated with Hadoop.

• Sqoop - Provides two way replication between Apache Hadoop and RDBMS. Supports snapshots and incremental updates.

• Kafka - distributed publish-subscribe messaging system. Hadoop is a consumer of Kafka.

• Storm - distributed computation based event-processing system. Often referred as real-time Hadoop. Storm cluster coordinates with Zookeeper.

• Others – Chukwa, Scribe, Samza,
Data Storage

Apache Hadoop –

• Framework used for multiple-node processing of data
  • Provides both distributed storage and distributed processing of very large data sets
• Scalable platform for processing large batches of data very fast; High degrees of parallelism
  • Master slave architecture
Apache Hadoop

Data Exchange
Sqoop
Flume
Log Control

Pig
Scripting
Hive
SQL
Mahout
ML
Oozie
Workflow

Zoo Keeper
Coordination

YARN/Map Reduce V2
Hadoop Distributed File System

…evolving*

AIOUG Sangam’16
Evolution of the Hadoop Ecosystem

Hadoop in 2010

**MapReduce**
One-stop shop for getting answers from your file system

**HDFS**
Hadoop Distributed File System, cheap shared storage for heterogeneous data

Hadoop Today

**MapReduce**

**Spark, Impala, HBase, etc...**
Streaming and in-memory

**Tez / Hive**
SQL on Hadoop

**Yarn**
Resource negotiator

**HDFS**
Distributed File System

Data scrubbing with Pig and Hive

Hive

- Data warehousing capability on top of Hadoop
- HiveQL provides familiarity SQL folks
- Uses MapReduce for execution
- Enable data mining on large volumes of data

Pig

- Dataflow scripting language
- Uses MapReduce for execution
- Pig interpreter submits the jobs to Hadoop cluster
Oracle Big Data SQL
Oracle Big Data SQL

- Powerful, high-performance SQL on Hadoop
- Full Oracle SQL capabilities on Hadoop
- SQL query processing local to Hadoop nodes
- Simple data integration of Hadoop and Oracle Database
  - Single SQL point-of-entry to access all data
  - Scalable joins between Hadoop and RDBMS data
- Optimized hardware
  - High-speed Infiniband network between Hadoop and Exadata
Transactional data-stores

- RDBMS do not scale with massive volumes of data
- NoSQL main characteristics is its non-adherence to relational database concepts of CODD
- Focus on scalability, performance, high availability
  - ACID properties are not always guaranteed
  - No joins, less complex, no constraints
Transactional data-stores

- Availability of data is more important than data consistency (BASE)
- Relations are addressed at application level
- Go by CAP theorem
  - Consistency, Availability, Partition tolerance
Transactional data-stores

• Key-value
  • Oracle NoSQL, DynamoDB, Voldemort, Apache Accumulo

• Document-based
  • MongoDB, CouchDB

• Column-based
  • Apache Cassandra, Apache Hbase

• Graph-based
  • Neo4J, InfoGrid

• Relational
  • Apache Kudu
Data Analytics

• Convergence layer where volume, velocity, and variety transform into Value

• Structured data analysis – Data mining, Inferential statistics

• Text Analysis – Natural Language Processing, Text mining, Opinion mining, categorization

• Web Analytics – Web mining, Web usage analysis

• Network Analytics – Social media based

• Mobile Analytics – location based mining

• Multi media analytics – event detection and prediction
## Build your Big Data team

<table>
<thead>
<tr>
<th>Role</th>
<th>Responsibilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Administrator</td>
<td>- Hadoop admins&lt;br&gt;- Information security&lt;br&gt;- DevOps</td>
</tr>
<tr>
<td>ETL Developer</td>
<td>- Implement ETL/ELT flow&lt;br&gt;- Sqoop, Flume, ETL tools, Stream processing</td>
</tr>
<tr>
<td>Data Architect</td>
<td>- Data modeling&lt;br&gt;- Hadoop&lt;br&gt;- ETL design&lt;br&gt;- Data analytics</td>
</tr>
<tr>
<td>Big Data Architect</td>
<td>- Develop core applications using NoSQL, Spark, MapReduce&lt;br&gt;- Data processing&lt;br&gt;- Data Visualization strategy</td>
</tr>
<tr>
<td>Data Scientist</td>
<td>- Statistical techniques&lt;br&gt;- Machine learning&lt;br&gt;- R, Python, Perl</td>
</tr>
</tbody>
</table>

AIOUG Sangam’16
Questions?
thank you