Big Data Approaches
Making Sense of Big Data

Ian Crosland
Jan 2016
Accelerate Big Data ROI

Even firms that are investing in Big Data are still struggling to get the most from it.

Qlik’s platform drives higher ROI by delivering big data in context with other data to ensure that Big Data stays relevant.
A strong culture of **Partnership**

**Broad range** of technology partnerships

**Dedicated** staff ensures continued focus

**Continuous evaluation** of new market entrants

**HADOOP**

**DATABASES**

**ACCELERATORS**

**DOMAIN**

**AND MORE...**
Apache Hadoop v2

- Hive (SQL)
- Pig (ETL)
- Mahout (ML)
- Giraph (Graph)
- HDFS2
- YARN (Cluster Resource Manager)
- MapReduce
- HBase
- Other Compute Engines (Tez, Spark, etc.)
- Hive 1.2.0, Tez, Pig, Cascading, etc.
- Mahout (ML)
- Hive 1.2.0
- Tez
- Pig
- Cascading
- Giraph (Graph)
- YARN (Cluster Resource Manager)
- HDFS2
Stinger Initiative

Hive on Tez
- YARN integration
- Distributed execution framework
- Eliminate extra map reads
- Dataflow model on DAG of nodes

Query Optimisations
- Vectorised query execution
- Filter at storage layer vs SQL engine
- SQL cost based optimiser

ORCFile format
- Higher compression
- Columnar
- Ideal for frequent fact filters

145 developers 44 companies

Source: http://hortonworks.com/labs/stinger/
**Impala**

- SQL Cost based optimisations
- Authentication, AD/Kerberos
- YARN integration
- In memory caching

**Parquet file format**
- Driven from Twitter use cases
- Columnar data storage
- Limits IO to data needed
- Space saving

**Metastore**
- Can be same DB as Hive metastore e.g. MySQL
- Query optimiser can use table/column stats
- Can use Hbase/HDFS with several file formats e.g. RCfile/Parquet

**Impala Roadmap**
- Additional SQL support
- S3 integration
- Nested data

Source: [http://blog.cloudera.com/blog/2014/08/whats-next-for-impala-focus-on-advanced-sql-functionality/](http://blog.cloudera.com/blog/2014/08/whats-next-for-impala-focus-on-advanced-sql-functionality/)
Apache Drill

Dynamic Schema Discovery
• Does not require schema/type spec to start query execution
• Leverage self describing data formats, e.g. Parquet, Avro, JSON and NoSQL DB
• Flexible data model built for complex/semi-structured data

Performance
• Distributed execution engine for query processing
• Columnar execution, avoids disk access for columns not in query
• Vectorisation allows CPU to operate on record batches
• Optimistic and pipelined query execution

Connect via ODBC

ODBC/JDBC

Drill SQL Query Layer & Execution Engine

Hive UDF’s

Files  HBase  Hive  Metastore  SerDes

Data sources
• Drill creates a virtual view in JSON
• Nested data support
• MAPR-FS, HDFS, H-Base
• Can use Hive Metastore
• JSON, Mongo DB/NoSQL
• Reuse Hive UDF’s

Source: https://www.mapr.com/products/apache-drill
Resilient Distributed Datasets
- In memory
- MR does not lend itself to interactive/ad-hoc queries
- Logical collection of data partitioned across machines
- Can reference external datasets

SparkSQL
- DataFrame distributed collection of data in named columns
- Supported on RDD’s, Parquet, JSON, Hive, JDBC sources
- YARN integration

Market
- Spark distributed with ALL hadoop distro’s
- Not just Big Data use cases

Source: https://databricks.com/spark/about
Qlik Big Data Methodologies

Different data volumes and complexities are best met using different methods.

Different methods ensure an optimized experience for the user for every situation.

**Methods can be combined** to meet different use cases.

Methods vary in deployment complexity.

**Data Volume**
- Size (rows)
- Dimensions (columns)
- Cardinality (uniqueness)

**App Complexity**
- Computational complexity such as set analysis
- Object density

Diagram:
- On Demand App Generation
- Direct Discovery
- In-Memory
- Chaining
- Segmentation
On Demand App Generation

A **shopping** cart approach to analytics

**Dimension** selection to generate filtered analytics

On demand data **slices**

Converting **Big Data** to small data analytics

Driven by business users **governed** by IT
Sample QlikView Process

**Selection App**
- Dimensions in list boxes
- Conditional show applied to a button to limit amount of data
- Action to invoke analysis document with parameters

**Index Technique**
- ASPX page invoked with selection criteria/user name
- EDX parameters passed separated in a list and pipe terminated

**Analysis App**
- Analysis app produced by EDX task
- Data slice limited by populating a WHERE clause with the parameters
Sample QlikView Server Process

1. **Selection App** with dimensional data is deployed to access point
2. **Business User** makes multiple data selections
3. **Governed** selections will drive the indexing of analysis app
4. **EDX** process indexes analysis app with latest data from source(s)
5. **Associative** analysis app available for business user

**Diagram Notes:**
- **User makes personalized selections from across many data sources**
- **Aggregated data and dimensions**
- **Selection app**
- **User selection criteria drives creation of analysis app**
- **Analysis app**
- **Data from multiple data sources is indexed into a template app based on the user selections**
- **Highly interactive user experience within a purpose built analysis app!**
Case Study - Telco

1. Selection App is populated with **dimensional data** on schedule

2. User selects dimensional criteria. After the **governed limit** is reached an ASPX page is invoked

3. QMS API and EDX indexes analysis app with the most recent data from the Teradata database with only the **data slice** relevant to the user

4. Analysis app deployed to access point with **user security**
Qlik Sense and Elastic Tweets Example

1. Tweets are populated into a Elastic DB via Logstash

2. User searches for Tweets stored in the Elastic DB from a custom web page

3. NodeJS container with Sense Proxy, Engine and Repository API’s and indexes the analysis app with the Tweets from the search stored in the Elastic DB. The data slice contains relevant for the user

4. Analysis app updated and published to a shared stream
### Authenticate

- Gets client certificate
- Generates XrfKey
- Send request for ticket
- Authenticates to Sense with ticket

### Generate and Publish App

- Access Index template app
- Copies template and renames with filter added to app name
- Replaces the $search_terms$ in the script with filter
- Generates app with data and publishes to everyone stream
Thank you