BI on Big Data

What are your options?

Tomer Shiran - @tshiran
Co-Founder & CEO, Dremio

Strata + Hadoop World London, June 3, 2016
BI on Hadoop: What are your Options

Dremio Company Background

- Stealth data analytics startup
- Founded in 2015
- Led by experts in Big Data and open source including the creators of Apache Arrow & Apache Parquet

Tomer Shiran
Founder & CEO

- Previously MapR (VP Product & employee #5), MapR; Microsoft; IBM Research
- Carnegie Mellon, Technion

Jacques Nadeau
Founder & CTO

- Recognized SQL & NoSQL expert
- Founder of Apache Arrow & Drill
- Previously Quigo (AOL); Offermatica (ADBE); aQuantive (MSFT)

Julien Le Dem
Architect

- Founder of Apache Parquet
- Apache Pig PMC Member
- Previously Twitter (Lead, Analytics Data Pipeline); Yahoo! (Architect)

Top Silicon Valley VCs

- Stealth data analytics startup
- Founded in 2015
- Led by experts in Big Data and open source including the creators of Apache Arrow & Apache Parquet

- Previously MapR (VP Product & employee #5), MapR; Microsoft; IBM Research
- Carnegie Mellon, Technion

- Recognized SQL & NoSQL expert
- Founder of Apache Arrow & Drill
- Previously Quigo (AOL); Offermatica (ADBE); aQuantive (MSFT)

- Founder of Apache Parquet
- Apache Pig PMC Member
- Previously Twitter (Lead, Analytics Data Pipeline); Yahoo! (Architect)
Recent changes to the BI landscape

Good ol’ Days

- Only a few databases (e.g. Oracle, Teradata, SQL Server)
- A few BI tools (MicroStrategy, Cognos)
- Everything worked with everything
- Things were easy!

Modern Reality

- Larger scale, less control and less structure
- Lots of databases!
- Data Lake, not database
- HDFS: It’s a file system, folks!
- NoSQL: Let’s put the schema in the application
- It can feel like the wild west!
Major Approaches to BI on Big Data?

ETL to RDBMS
- “Make the new world look like the old world!”
- Load a transformed set of data into relational database

Monolithic (all-in-one) solutions
- Use BI tools that connect directly to Big Data

SQL-on-Big-Data
- Connect BI tools to a query engine sitting on top of Big Data
- Three main sub-categories
  - Native SQL
  - Batch SQL
  - OLAP Cubes
So how do we bring BI to Big Data?

ETL tool → RDBMS

BI options

Monolithic tool with built-in BI

Big Data

Monolithic All-in-one Solutions

SQL-on-Big-Data

BI options

SQL Engine

Big Data
ETL to RDBMS: Introduction

- ETL (Extract, Transform, and Load) a subset of the data into a relational database
  - Oracle, PostgreSQL, Teradata, Redshift, Vertica, ...

- Connect any desired BI tool to the RDBMS
  - Tableau, Qlik, ...

- Two options:
  - Commercial tools (Informatica, Talend, Pentaho, ...)
  - Custom development, scripts, etc.
ETL to RDBMS: Example

- Load web server logs from HDFS into RDBMS
- ETL software: Pentaho Data Integration (aka ‘Kettle’)
- RDBMS: MySQL
ETL to RDBMS: Pros and Cons

Pros

• Relational databases and their BI integrations are very mature
• Use your favorite tools
  o Tableau, Excel, R, …

Cons

• Traditional ETL tools don’t work well with modern data
  o Changing schemas, complex or semi-structured data, …
  o Hand-coded scripts are a common substitute
• Data freshness
  o How often do you replicate/synchronize?
• Data resolution
  o Can’t store all the raw data in the RDBMS (due to scalability and/or cost)
  o Need to sample, aggregate or time-constrain the data

…and really, who wants to ETL?
Monolithic (or All-in-One) Solutions: Introduction

- Single piece of software on top of Big Data
- Performs both data visualization (BI) and execution
- Utilize sampling or manual pre-aggregation to reduce the data volume that the user is interacting with
- Examples:
  - Datameer
  - Platfora
  - Zoomdata
Platfora Architecture Overview

- Constructs aggregates that are loaded into an external database
  - Aggregates provide fast visualizations
  - Aggregations must be created before consumption
Datameer Architecture Overview

- Users interact with samples of the data in an Excel-like interface
- Finished designs use the whole dataset
- Query router determines execution engine based on data size

Hadoop Cluster

HDFS

Datameer Nodes

Single Node Custom Execution

Query Router

Sampling

Tez

MapReduce
• Queries on historical (ie, non-streaming) data are split into many sampling queries
• This sampling provides a view of the data that converges toward an accurate picture
  - But adds load on the data source...
• Can handle streaming data sources
Monolithic Solutions: Pros and Cons

Pros
- Only one tool to learn and operate
- Easier than building and maintain ETL-to-RDBMS pipeline
- Integrated data preparation in some solutions

Cons
- Can’t analyze the raw data
  - Rely on aggregation or sampling before primary analysis
- Can’t use your existing BI or analytics tools (Tableau, Qlik, R, …)
- Can’t run arbitrary SQL queries
SQL-on-Big-Data: Introduction

• SQL queries against Big Data
  o Hadoop
  o NoSQL
    • MongoDB, HBase, ...
  o Cloud Storage
    • S3, Azure Data Lake, GCS, ...

• Use your existing BI tools
  o Leverage standard ODBC/JDBC drivers

Tableau, Qlik, R, ...
SQL Engine
Hadoop & NoSQL
SQL-on-Big-Data: Introduction

Three major design philosophies:

- Native SQL
- Batch & Data Science SQL
- OLAP Cubes on Hadoop
Native SQL

- Apache Drill
  - Queries Hadoop, RDBMS, Files, NoSQL, Cloud (S3)
  - Based on Apache Arrow
  - Columnar in-memory execution

- Apache Impala (incubating)
  - Utilizes the Hive metastore
  - Focused on data in HDFS

- Presto
  - Queries Hadoop, RDBMS, Files, NoSQL, Cloud (S3)
Native SQL: Pros and Cons

Pros

• Highest performance for Big Data workloads
• Connect to Hadoop and also NoSQL systems
• Make Hadoop “look like a database”

Cons

• Queries may still be too slow for interactive analysis on many TB/PB
• Can’t defeat physics
Batch & Data Science SQL

• Hive
  o Enables SQL queries to be translated to MapReduce/Tez
  o Most commonly used for batch processing and ETL workloads

• Spark SQL
  o Provides a way to deliver SQL queries in Spark programs (Scala/Java/Python)
  o Excellent interleaving with data science work
Batch & Data Science SQL: Pros and Cons

Pros
- Potentially simpler deployment (no daemons)
  - New YARN job (MapReduce/Spark) for each query
- Check-pointing support enables very long-running queries
  - Days to weeks (ETL work)
- Works well in tandem with machine learning (Spark)

Cons
- Latency prohibitive for interactive analytics
  - Tableau, Qlik Sense, …
- Slower than native SQL engines
OLAP Cubes on Hadoop

- **Kylin**
  - Hadoop-only
  - Stores OLAP cubes in HBase
  - Queries fail if not satisfied by cubes
  - Open source

- **AtScale**
  - Hadoop-only
  - Leverages external SQL engine
    - Hive, Impala, SparkSQL
  - Collaborative cube creation
  - Closed source
OLAP Cubes on Hadoop: Pros and Cons

Pros
- Fast queries on pre-aggregated data
- Can use SQL and MDX tools

Cons
- Explicit cube definition/modeling phase
  - Not “self-service”
  - Frequent updates required due to dependency on business logic
- Aggregation create and maintenance can be long (and large)
- User connects to and interacts with the cube
  - Can’t interact with the raw data
## SQL-on-Big-Data: Solution Comparison

<table>
<thead>
<tr>
<th></th>
<th>Native SQL</th>
<th>Batch &amp; DS SQL</th>
<th>OLAP Cubes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Technologies</strong></td>
<td>Drill, Impala, Presto</td>
<td>Hive, Spark SQL</td>
<td>Kylin, AtScale</td>
</tr>
<tr>
<td><strong>Connectivity</strong></td>
<td>SQL and NoSQL</td>
<td>SQL and NoSQL</td>
<td>Hadoop-only</td>
</tr>
<tr>
<td><strong>Primary Use Case</strong></td>
<td>Interactive</td>
<td>ETL or data-science focused</td>
<td>Constrained Interactive</td>
</tr>
<tr>
<td><strong>Query Capability</strong></td>
<td>Raw data</td>
<td>Raw data</td>
<td>Aggregated data</td>
</tr>
<tr>
<td><strong>Deployment Model</strong></td>
<td>New daemons collocated with existing services</td>
<td>New MapReduce and/or Spark job for each query</td>
<td>Varies</td>
</tr>
</tbody>
</table>

**Primary Use Case**
- Interactive: Raw data
- ETL or data-science focused: Aggregated data
- Constrained Interactive: Varies
SQL-on-Big-Data: General Pros and Cons

Pros

• Continue using your favorite BI tools and SQL-based clients
  o Tableau, Qlik, Power BI, Excel, R, SAS, …
• Technical analysts can write custom SQL queries

Cons

• Another layer in your data stack
• May need to pre-aggregate the data depending on your scale
• Need a separate data preparation tool (or custom scripts)
Deciding what is right for you?
**BI on Big data: Heuristic**

1. **Do you already have a favorite BI Tool**
   - **Yes**: Does your schema change frequently?
     - **No**: Is your working data relatively small & static?
       - **Yes**: Do you have very predictable analysis needs?
         - **No**: Do you need to query NoSQL?
           - **Yes**: OLAP Cubes on Hadoop
           - **No**: Are you focused on interactive BI?
             - **No**: Do you want to combine ML with SQL?
               - **Yes**: Native SQL
               - **No**: SparkSQL
             - **Yes**: Datameer
             - **No**: Hive
           - **Yes**: Datameer
         - **Yes**: Datameer
       - **Yes**: Platfora
     - **Yes**: ETL to RDBMS
   - **No**: Do you want to be able to write SQL?
     - **Yes**: Is External Cluster Okay?
       - **No**: Datameer
       - **Yes**: Zoomdata
     - **No**: Datameer

2. **Is External Cluster Okay?**
   - **Yes**: ETL to RDBMS
   - **No**: Platfora

3. **Do you like Excel Metaphor?**
   - **No**: Datameer
   - **Yes**: Zoomdata

4. **Monolithic/All-in-one Solutions**
   - Platfora
   - Datameer
   - Zoomdata
   - Hive
   - SparkSQL
   - Native SQL
   - OLAP Cubes on Hadoop
Q&A

Tomer Shiran
tshiran@dremio.com
@tshiran

Reach out to learn what we’re up to at Dremio (or to join the private beta…)

BI on Hadoop: What are your Options