BI on Big Data

Keys to Success for BI on Hadoop

Josh Klahr
VP of Product
josh@atscale.com
@jklahr
HELLO!
My name is Josh

A long time ago
When Yahoo! was relevant
Hadoop wars...
What I learned along the way

Data movement is the source of all evil

Business users love OLAP interfaces

You must have standard definitions

You can’t have just one BI front end; also, Excel...

Data movement is the source of all evil

Slow queries = no users

Oh, security...
What BI tools do you use? Survey says...

Tableau: 51% (2016) vs 59% (2015)
Excel: 37% (2016) vs 45% (2015)
SAP/Business Objects: 23% (2016) vs 36% (2015)
Microstrategy: 15% (2016) vs 19% (2015)
IBM Cognos: 16% (2016) vs 19% (2015)
Microsoft PowerBI: 16% (2016) vs 12% (2016)
Qliktech: 12% (2016) vs 12% (2016)
Domo: 4% (2016) vs 12% (2016)

Want more? Take the assessment and get our maturity assessment @ www.atscale.com/survey
**Bi on Hadoop with data movement**

1. New Data Type on Cluster
2. Hadoop Dev Team Extracts new Element
3. Hadoop Dev Builds Pre-Aggregate Table
4. ETL Team Updates Data Mart Load Scripts
5. DB Dev Team Updates Base Schema
6. DB Dev Team Creates Summary Tables
7. BI Server Team Updates Metadata
This approach is no fun

- Increased cost of storage
- Expensive development cycles
- Lack of agility
- No shared definitions

Everyone is unhappy (and out of money)
### BI on Hadoop: NOT SQL on Hadoop

<table>
<thead>
<tr>
<th>Business Use Case</th>
<th>Calculation Engine Requirement</th>
<th>Industry Vertical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Per-Member-Per-Month</td>
<td>Multi-fact Metrics</td>
<td>Health, Insurance</td>
</tr>
<tr>
<td>Counting Unique Customers</td>
<td>Semi-additive Metrics</td>
<td>Media</td>
</tr>
<tr>
<td>Store vs. Region Performance</td>
<td>Multi-level Metrics</td>
<td>Retail</td>
</tr>
<tr>
<td>Credit/Claims Fraud Detection</td>
<td>Atomic Drill-through</td>
<td>Financial Services</td>
</tr>
<tr>
<td>Period to Period Comparisons</td>
<td>Parallel Period</td>
<td>Retail</td>
</tr>
<tr>
<td>Currency Fluctuations</td>
<td>Metrical Attributes</td>
<td>Financial Services</td>
</tr>
</tbody>
</table>
Different definitions = poor decisions

USE THE CRS DATABASE TO SIZE THE MARKET.

THAT DATA IS WRONG.

THEN USE THE SIBS DATABASE.

THAT DATA IS ALSO WRONG.

CAN YOU AVERAGE THEM?

SURE. I CAN MULTIPLY THEM TOO.
Querying billions of rows of data… never fast enough
Benchmarks!

TECH TALK: BI Performance Benchmarks with Google BigQuery
Posted by Joshua Klahr on Apr 6, 2017

TECH TALK: BI-on-Hadoop Engine Wars Continue...Everybody Wins
Posted by Joshua Klahr on Oct 18, 2016

TECH TALK: SQL-on-Hadoop Benchmark: A Bit of a Tortoise and Hare Story
Posted by Trystan Leftwich on Feb 24, 2016

More @ http://info.atscale.com/benchmark
Security policies are useless if they don’t cover the stack
It's not all doom and gloom
Yes! You can have a modern framework for BI on Big Data

1. Treat your data as a shared asset
2. Provide the right interfaces for consumption
3. Provide information through data stewardship
4. Ensure a common vocabulary for analysis
5. Deploy comprehensive security and access controls
6. Avoid data movement whenever possible

Leverage a platform that provides a true OLAP experience

Rich metadata model

Built-in business analytics

Open data interfaces

Automatic Big Data performance optimization
Use your Big Data cluster to support BI queries too!

20X to 100X faster query times
Provide a schema-on-demand experience

Support data in its natural format

Create attributes on the fly

Enable analytics agility
Ensure you are leveraging the Hadoop security ecosystem

Delegated authorization is a MUST

Don’t pursue strategies that require multiple policy layers

Enable secure multi-user access to shared data assets
"We are moving our entire data warehouse to run on **BigQuery**. But the missing piece of the puzzle for us is a **live OLAP interface** to the data."

**Why Atscale?**

- Delivers queries under 5 seconds, **hundreds of users**
- Supports ad hoc Excel pivot table queries
- Handles retail calendars, multi-level metrics

**What’s the payoff?**

- Deliver 3x the historical data w/ lower latency
- Eliminate $100M in Teradata cost in 18 months
- Unblock migration to Google Cloud
“Support thousands of BI queries per day to allow suppliers to manage store inventory and sales for 76 billion rows of SKU level data in the Cloud”.

Why Atscale?
- Delivers queries under 5 seconds
- Supports Azure with high concurrency
- Handles retail calendars, multi-level metrics

What’s the payoff?
- Eliminate 4 copies of data
- Reduce Teradata cost
- Unblock migration to Azure Cloud
Deploy an on-cluster, scale-out solution
Leverage data platform security
Look for smart query optimization
Ensure support for true OLAP business models
Pursue a shared-semantic layer for all tools
Thank You!

Questions? Comments? Discussion?

Josh Klahr
VP of Product
@jklahr