

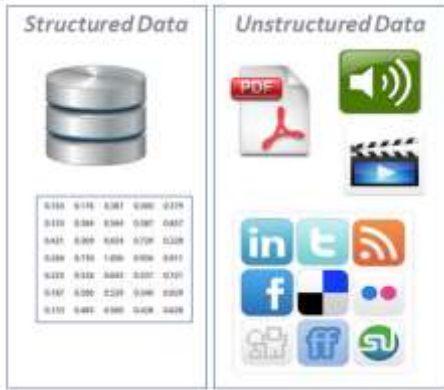


BigDataRevealed

Investigates and Discovers Data That Positively or
Negatively Impacts Your Business Performance



Top Challenges for Managing Big Data



Data is highly fragmented, in databases, emails, PDF's, spreadsheets and more.



Exploding real-time data from social media and IoT sources, arriving in high data volume



U.S. and European banks paid nearly \$65 billion in penalties and fines, about 40% greater than 2013, the previous high, according to the Boston Consulting Group.



IoT – Collecting & Analyzing real-time data such as machine health check logs



Growing concerns about release of personal information, security & privacy of all data



Recent World Events Supporting BDR Business Value



EU DPA Regulation

The regulation returns control over citizens' personal data to citizens



F.C.C. Fines AT&T **\$25 Million** for Privacy Breach

NY Times, April 2015



Good news! Big banks only have **\$65 billion in legal fines** left to pay.

Yahoo Finance, August 26, 2015



It is estimated that poor data quality costs US companies **\$600 billion per year**

TechRepublic, December 21, 2015



In the United States, it is reported that by 2018 there will be more than 490,000 data science positions available, but only 200,000 qualified people to fill the roles.

Datanami, January 22, 2016



Recent World Events Supporting BDR Business Value



HIPAA Violation

Lahey Hospital and Medical Center in Burlington, Mass., agreed to pay \$850,000 to settle potential HIPAA violations

On September 2, 2015 The HHS Office of Civil Rights (OCR) issued a press release



HIPAA Violation

University of Washington Medicine in Seattle agreed to pay \$750,000 to settle violation allegations.

On September 2, 2015 The HHS Office of Civil Rights (OCR) issued a press release



A 2011 report from McKinsey Global Institute predicted by 2018, the U.S. could face a shortage of 140,000 to 190,000 qualified data analysts, as well as 1.5 million managers who know how to use big data to make decisions.

Worcester Business Journal, January 4, 2016

BDR Support for the Process and Analyze Phase of the Big Data Life Cycle



Process

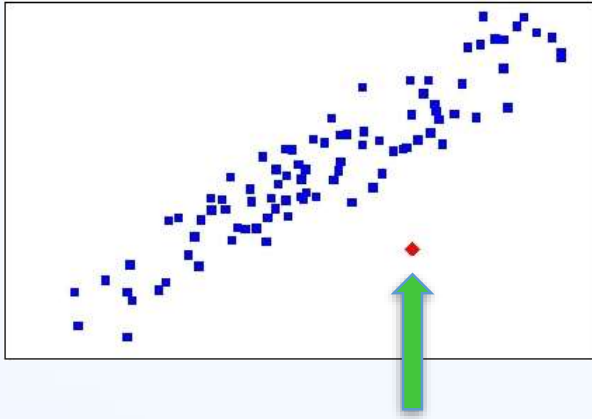
- Classification
- Cataloging
- Meta Data Roadmap
- Accuracy
- Streaming Data
- Pre-Built Database Connectors

Analyze

- Outlier Discovery
- Correlation Coefficients
- Provide Relevant Data to BI and Analytics



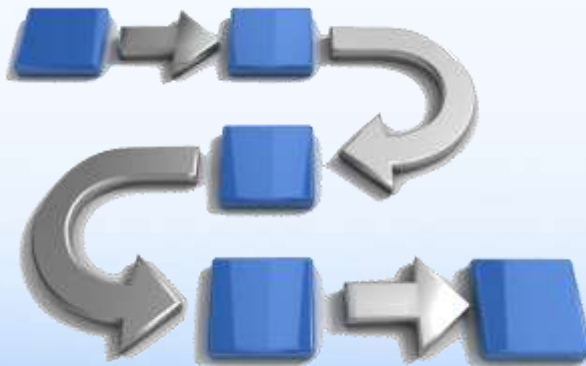
Investigate and Discover Data



Discover outliers that impact business performance and Compliance Violations



Data Scientists are expected to find correlation coefficients, outliers and other data anomalies
"According to a report from Experian, the average company estimates 27 percent of its revenue is wasted due to inaccurate or incomplete data"



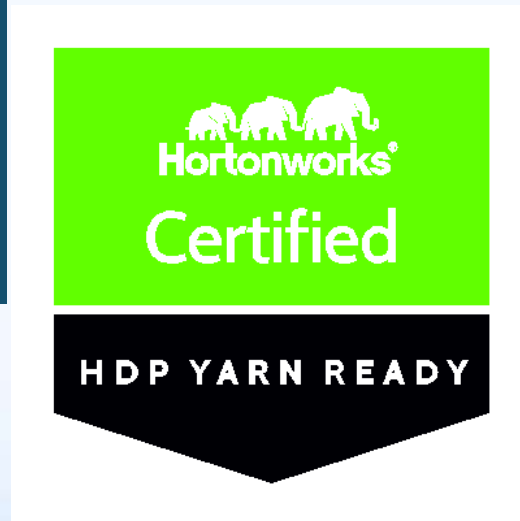
Fortune 1000 companies need solutions that plug into existing business processes



Unique Value

- 1.** Big Data Revealed Runs On The Hadoop Native Framework Leverages Existing Investment In Technology
- 2.** Data Discovery, Compliance Validation, Anomalies, Outlier Detection/Alerts, and User Definable Discovery
- 3.** Read/Source Data Directly: HDFS, Teradata, Oracle, DB2, MySQL, PDF, DOCX, HTML, Excel and more
- 4.** A Complete Solution - Repeatable, Collaborative, And Extensible
- 5.** Process Static HDFS, RDBMS Data As Well As Live Streaming Data Feeds With Real-time Discovery
- 6.** Run With BDR GUI Or Use Callable Modules Within Current Production Processes And ETL/BI Processes

Certifications





Cloudera Support

cloudera[®]
IMPALA

cloudera[®]
navigator

A P A C H E
HBASE



Spark





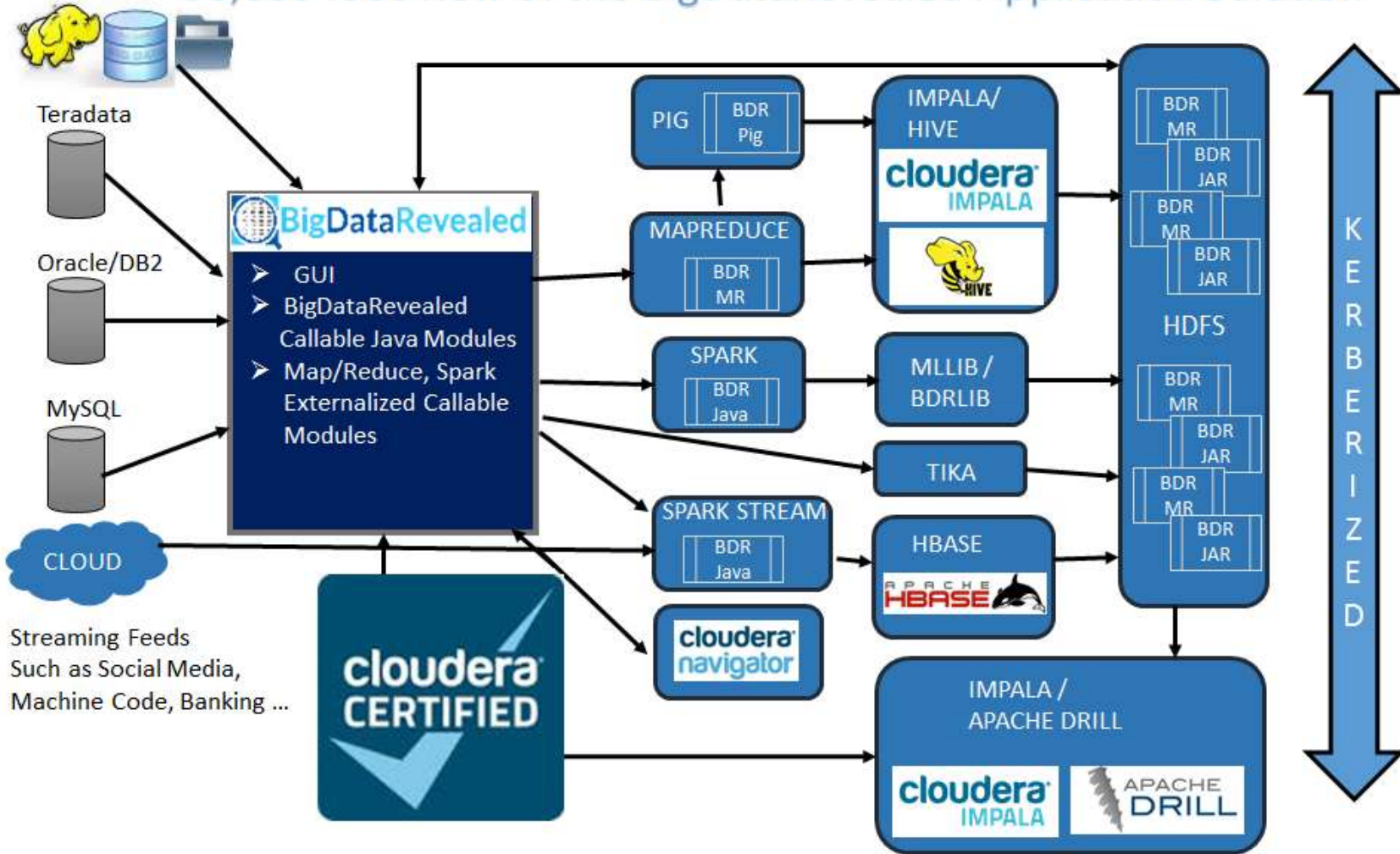
Kerberos Support

- **Kerberos** /'kərbərəs/ is a computer network authentication protocol which works on the basis of 'tickets' to allow nodes communicating over a non-secure network to prove their identity to one another in a secure manner.





50,000 foot view of the BigDataRevealed Application Solution





Architecture Overview

- Installs on any cluster client
 - *Database jobs require distributing drivers to the cluster
- Web front end runs on JBOSS 7.1.1
 - Defaults to port 8282
 - Requires MySQL for application support tables
- Utilizes Map Reduce and Spark via Yarn
 - Most jobs store in Hive/HBase
- Leverages Cloudera Impala for performant queries
- Leverages Cloudera Navigator for enhanced metadata

Architecture Legend

- Dark Blue Components are part of BigDataRevealed



BigDataRevealed REST API

- Light Blue Components are part of the Cloudera ecosystem



Cloudera
Navigator



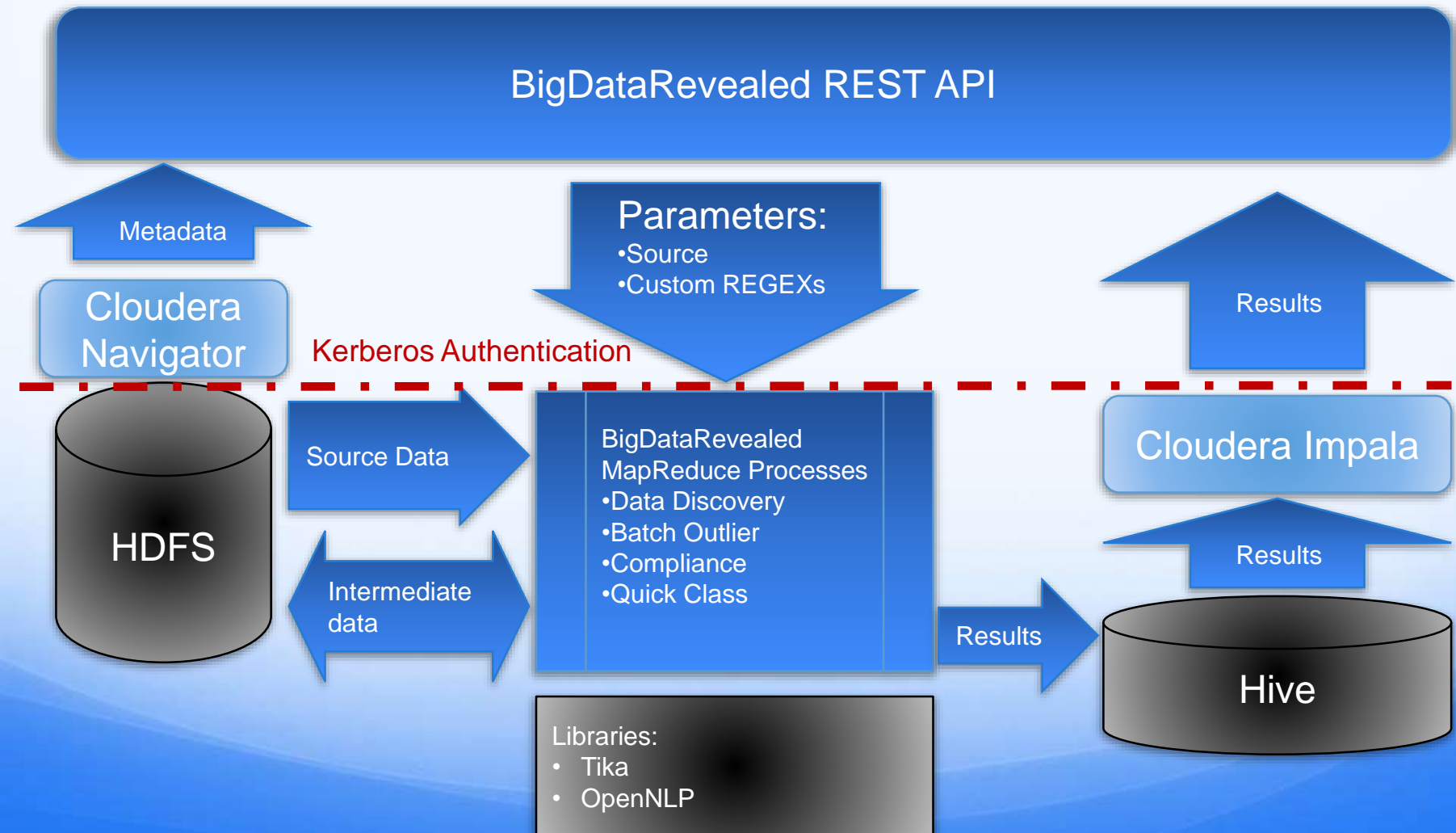
Cloudera Impala

- Black/Grey Components are part of the larger open source ecosystem



Hive

Anatomy of a BDR Map Reduce Job

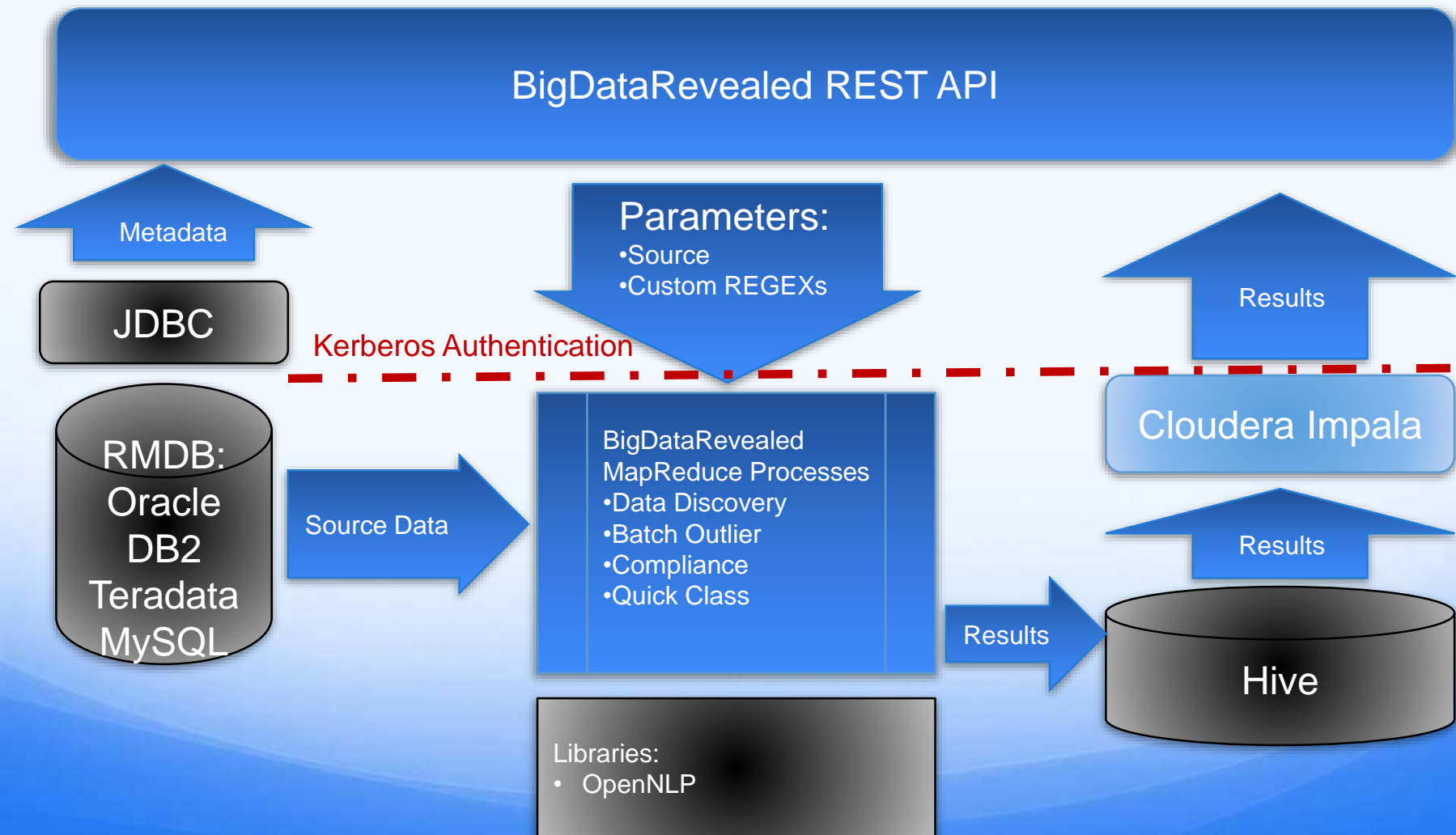




Anatomy of a BDR Map Reduce Job

- UI utilizes Cloudera Navigator and webHDFS to gather file system information
- REST API initiates the map reduce job with user selected parameters
- If the files are binary, they are processed through Tika with intermediate results stored back to HDFS
- Batch Process executes and stores results in Hive
- Results are checked against user watch conditions and notifications sent
- UI leverages Cloudera Impala to provide a performant view into result data

Anatomy of a BDR Database Job

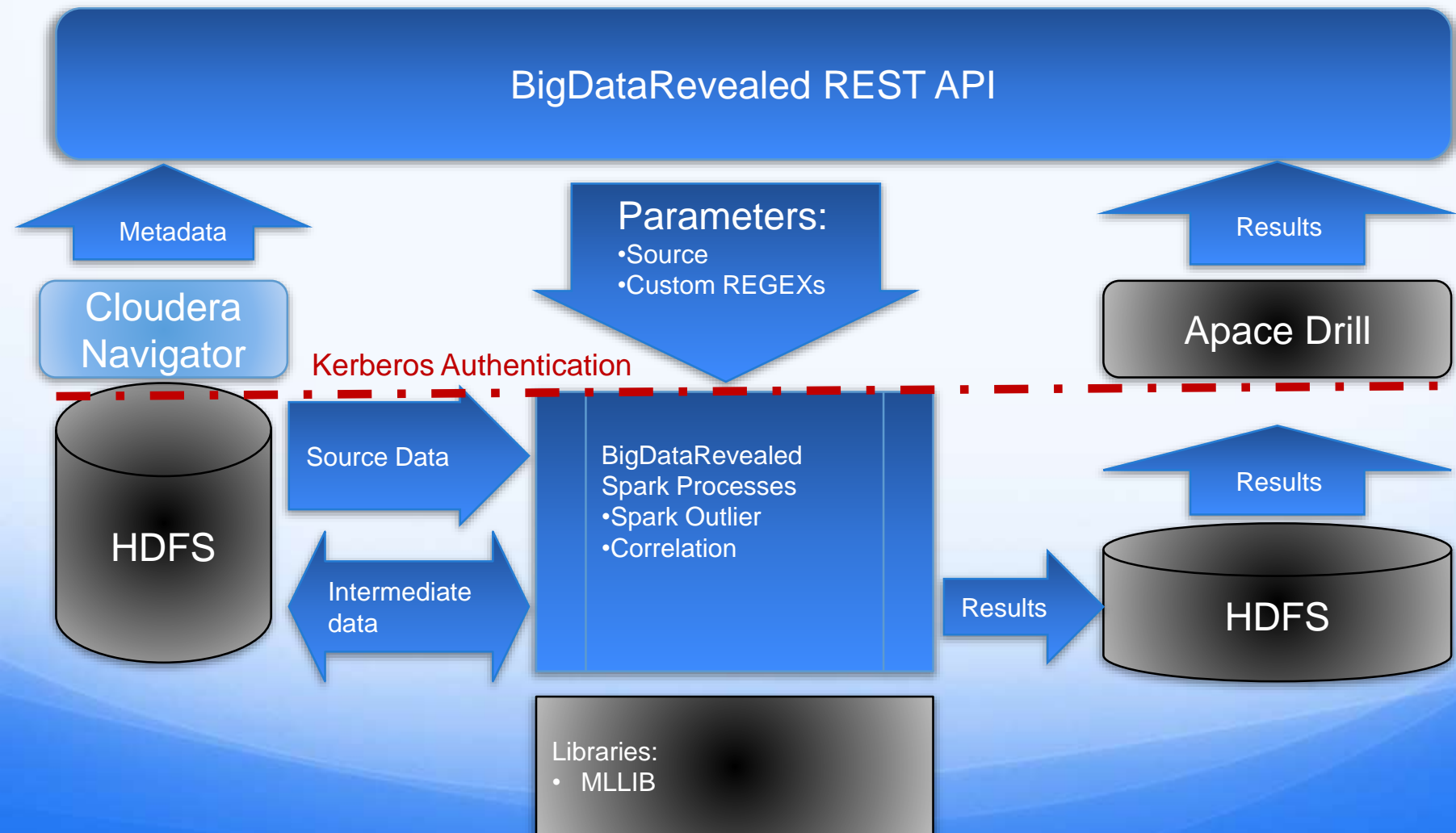




Anatomy of a BDR Database Job

- UI utilizes JDBC to gather database metadata
- REST API initiates the map reduce job with user selected parameters
- Batch Process executes and stores results in Hive
- Results are checked against user watch conditions and notifications sent
- UI leverages Cloudera Impala to provide a performant view into result data

Anatomy of a BDR Spark Job

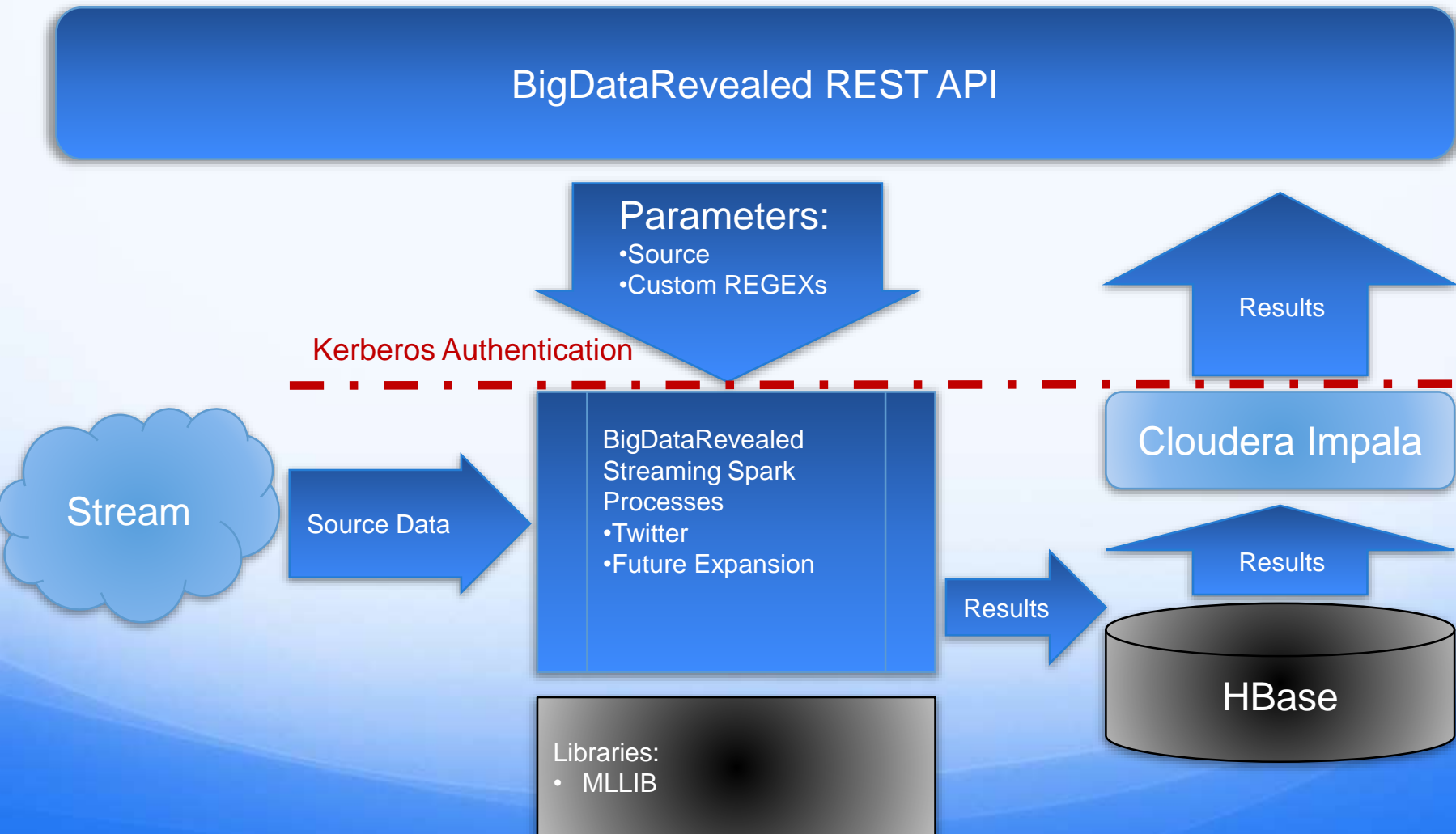




Anatomy of a BDR Spark Job

- UI utilizes Cloudera Navigator and webHDFS to gather file system information
- REST API initiates the spark job with user selected parameters
- Spark Process executes and stores results in HDFS
- Results are checked against user watch conditions and notifications sent
- UI leverages Apache Drill to view HDFS file contents

Anatomy of a BDR Streaming Job





Anatomy of a BDR Streaming Job

- User selects from a list of Admin entered data streams
- REST API initiates the spark job with user selected parameters
- Spark executes and stores results in HBase
- UI leverages Cloudera Impala to provide a view into result data



BigDataRevealed

Investigates and Discovers Data That Positively or
Negatively Impacts Your Business Performance

BigDataRevealed Contacts

Steven Meister
BigDataRevealed
steven.meister@bigdatarevealed.com
847-791-7838