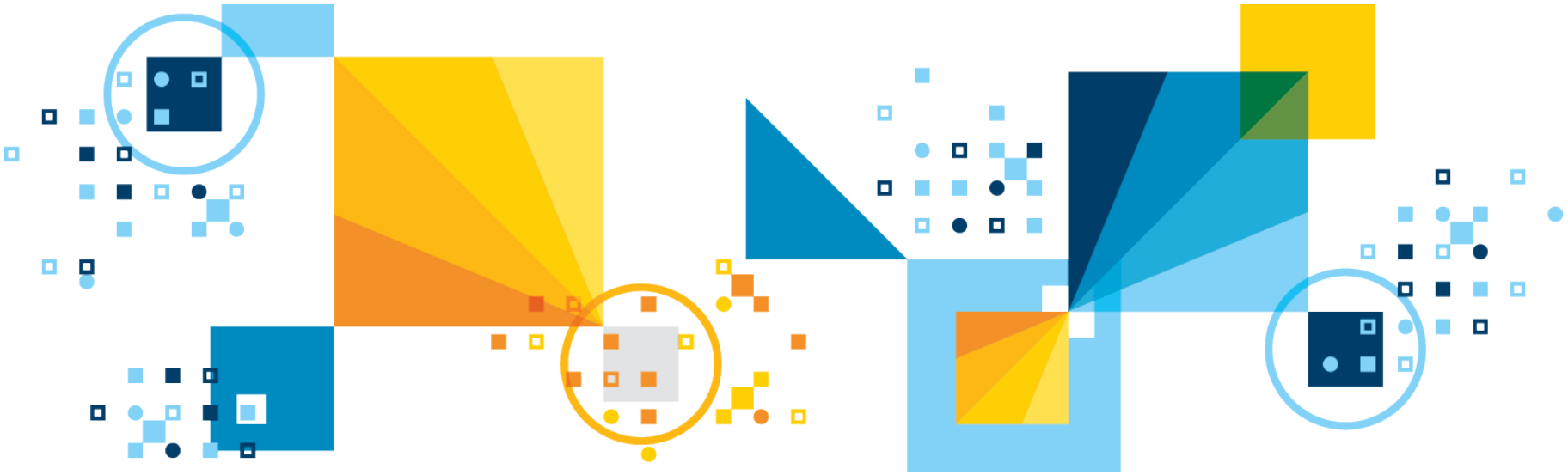
Francisco Cano
francisco.cano@ibm.com
June 2017

IBM BigInsights:
Bringing you big value from Big Data



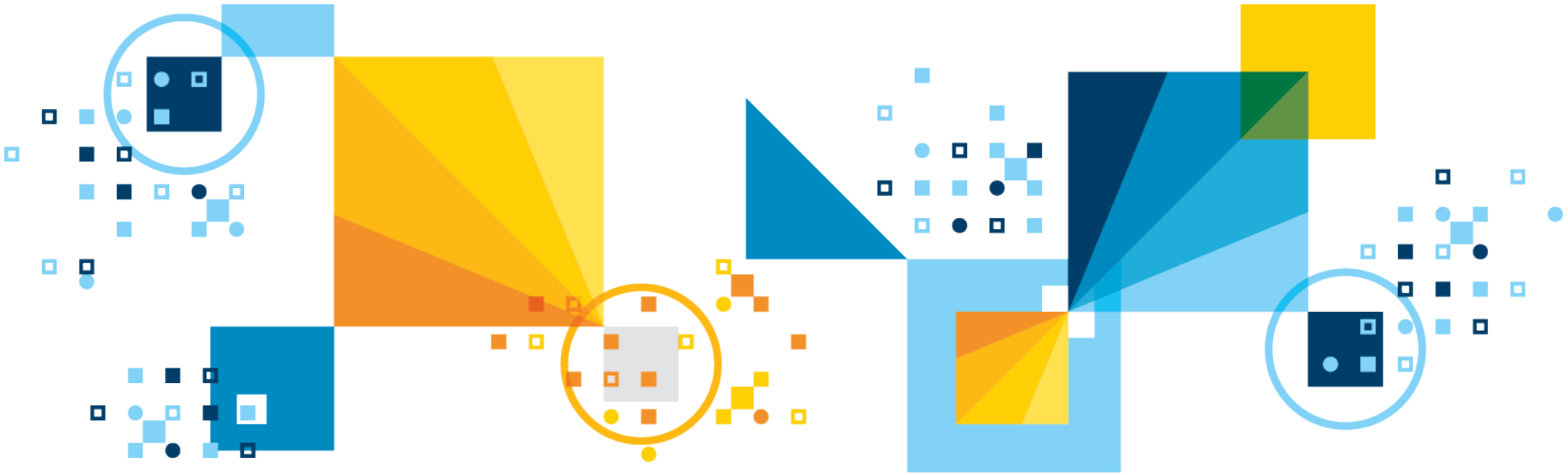
Agenda

- The big picture about Big Data
- IBM's approach
- How IBM can help you get off to a quick start
- Labs!!!!

big data big

"With great power comes great responsibility"

The Big Picture about Big Data



What is Big Data?

Definition

The realization of **greater business intelligence by storing, processing, and analyzing large volumes of structured, unstructured and semi-structured data** that was previously ignored **due to the limitations** of traditional data management technologies.

Multitudes of sources

- Existing databases or warehouses
- RFID readers
- Scanners
- Machine sensors
- Web page log files

Example:

Traditional data

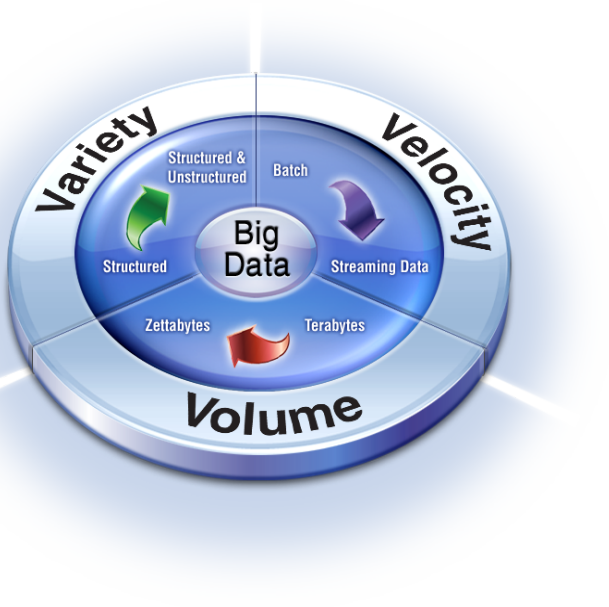
- Customer information and order history stored in RDBMS and used in online transactions.

Big Data

- 300 TBs of web log files from millions of visits to your companies website over the last several years.

Big Data presents big opportunities

Extract insight from a high volume, variety and velocity of data in a timely and cost-effective manner

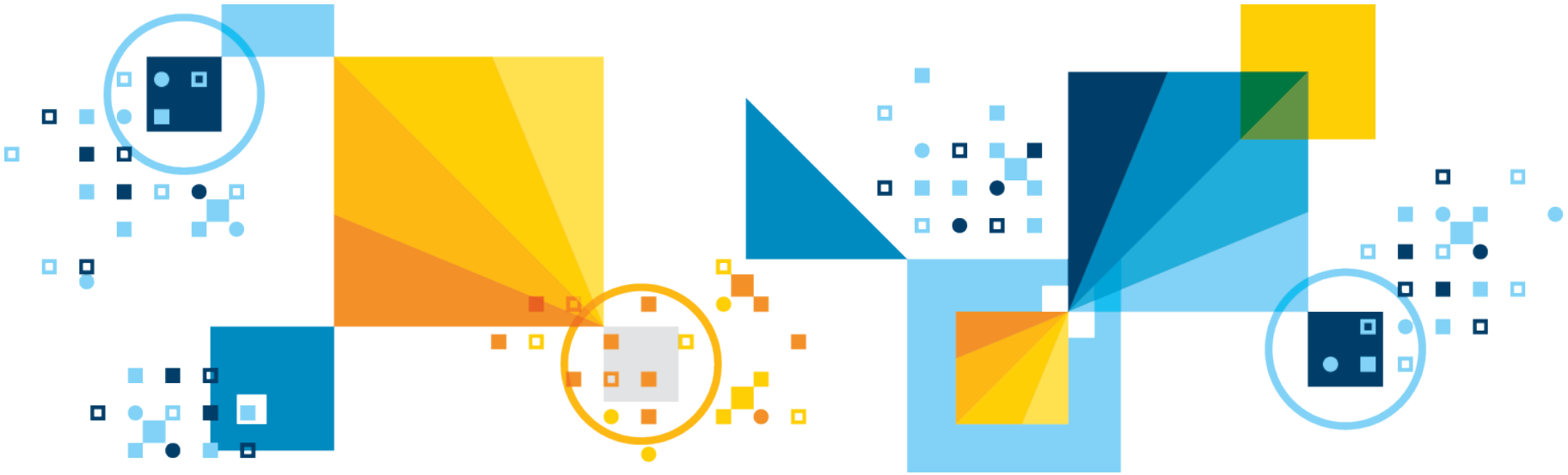


Variety: Manage and benefit from diverse data types and data structures

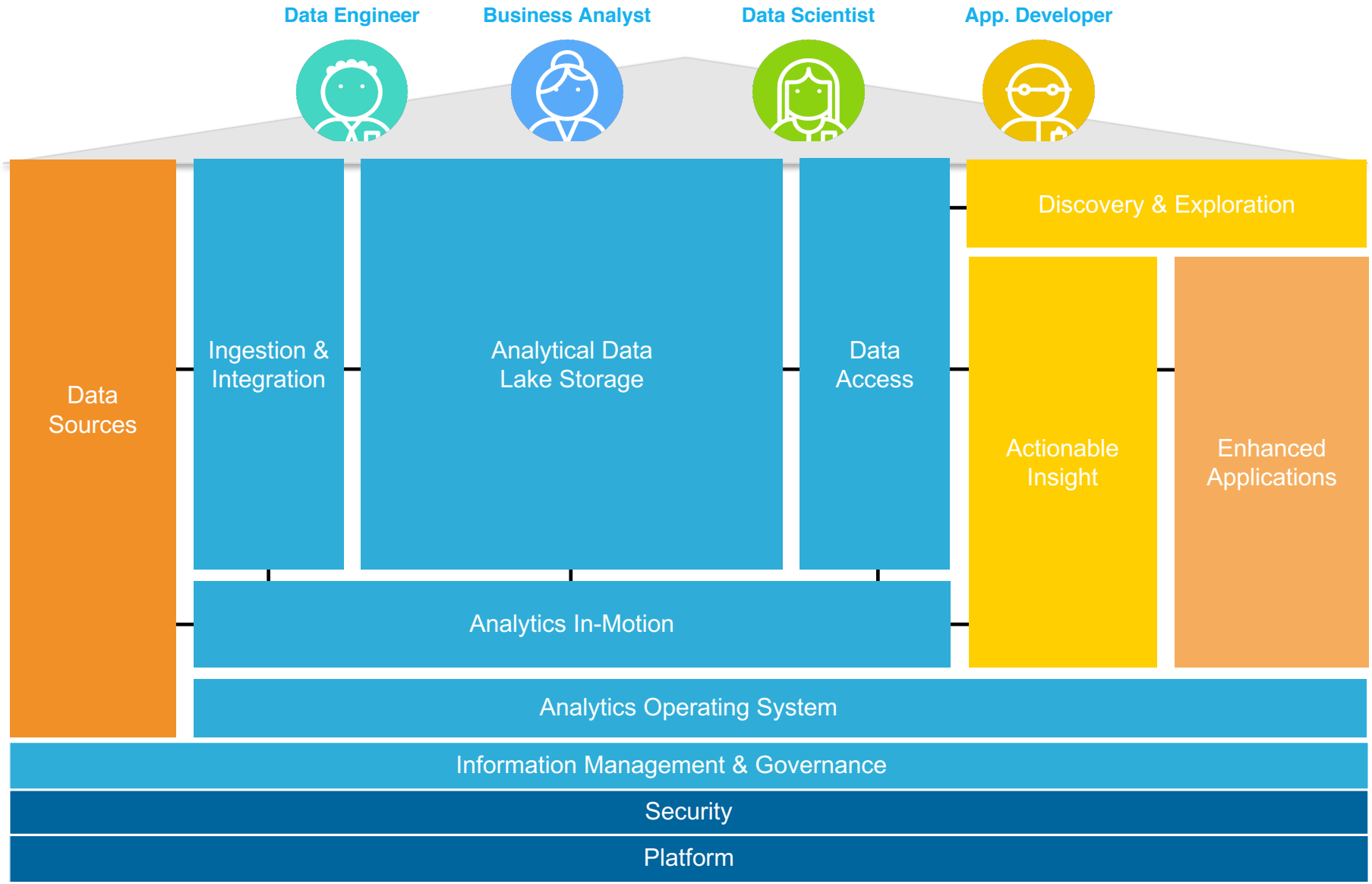
Velocity: Analyze streaming data and large volumes of persistent data

Volume: Scale from terabytes to zettabytes

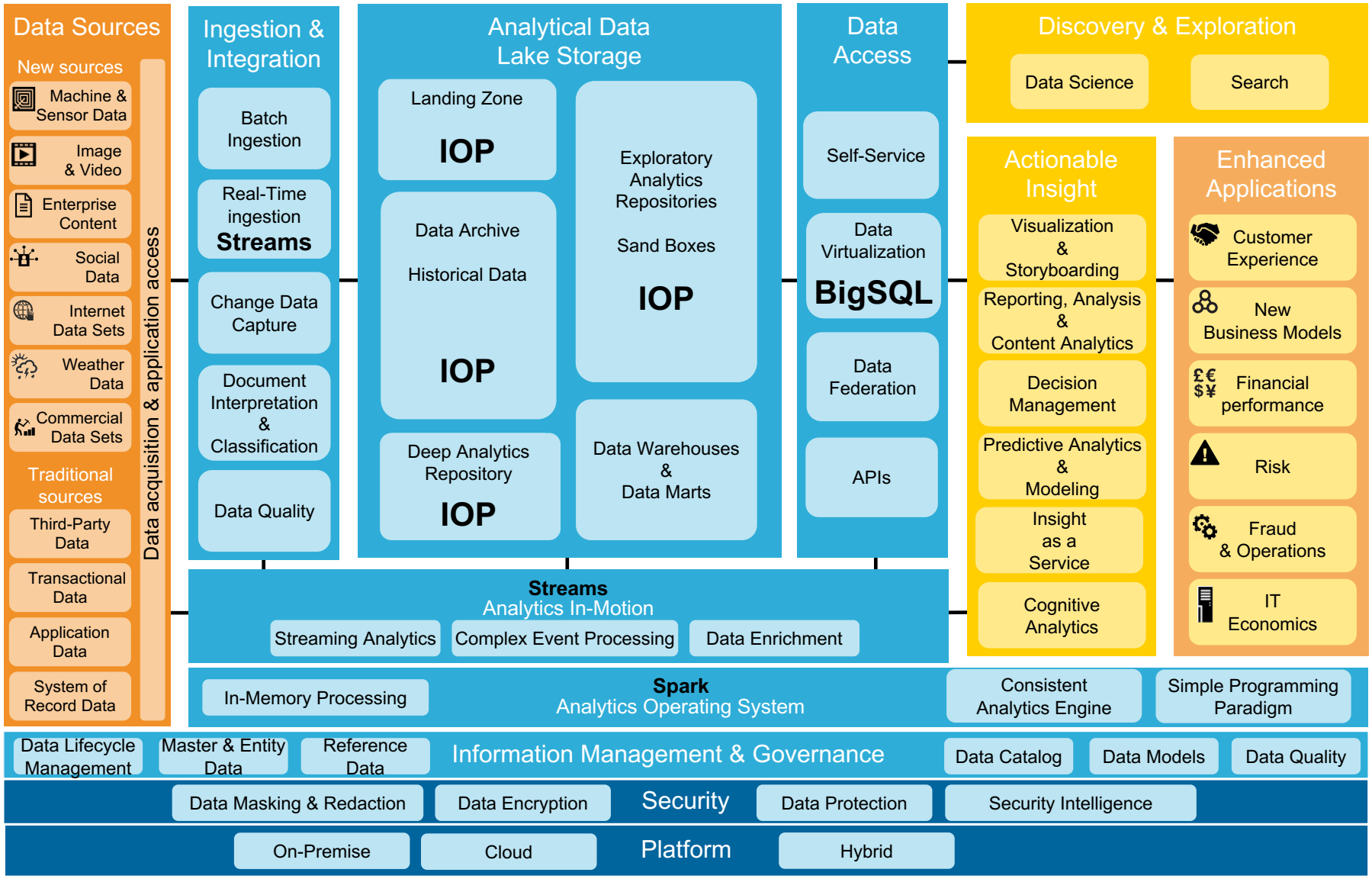
IBM's approach



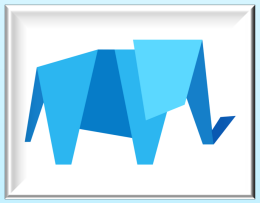
IBM Analytics Reference Architecture: Components & Personas



IBM Analytics Reference Architecture: Capabilities & Big Insights



Overview of BigInsights



IBM-specific BigInsights features

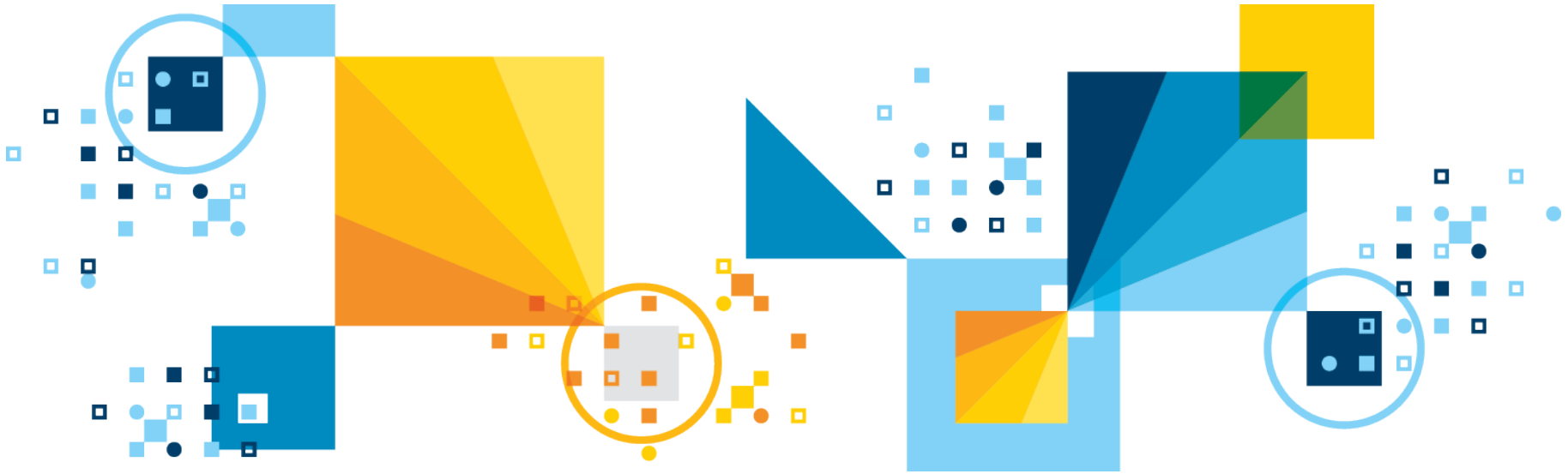
Big SQL (industry standard SQL)
Text analytics
BigSheets (spreadsheet-style tool)

IBM Streams, Cognos (limited use licenses)

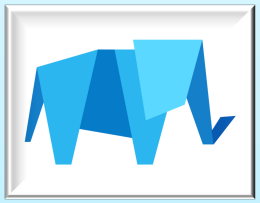
IBM Open Platform

100% open source platform compliant with ODPi
Apache Hadoop ecosystem
Apache Spark ecosystem

A Closer Look at IBM BigInsights . . . .



Overview of BigInsights



IBM-specific BigInsights features

Big SQL (industry standard SQL)
Text analytics
BigSheets (spreadsheet-style tool)

IBM Streams, Cognos (limited use licenses)

IBM Open Platform

100% open source platform compliant with ODPi
Apache Hadoop ecosystem
Apache Spark ecosystem

IBM Open Platform Harmonize on ODPi

Doubled
member companies

35
technical maintainers

- Runtime certification released – a technology sandbox and test suites
- **IBM Open Platform cer cation expected as part of v4.2**

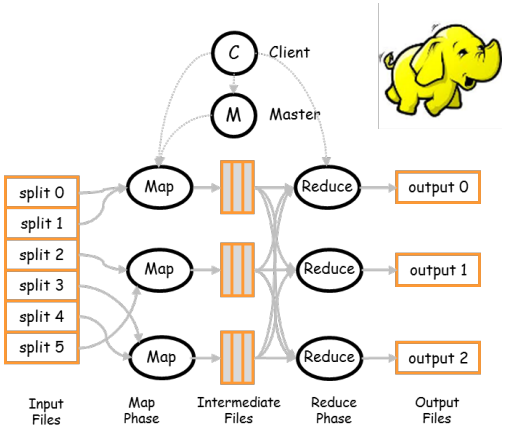
cloudera®

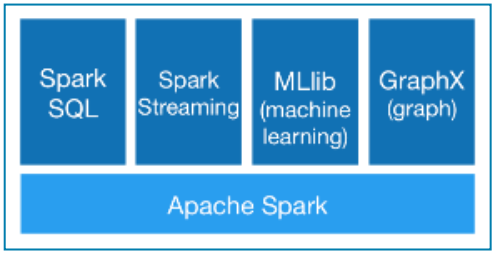
✓ VM sandbox for simple experimentation
✓ Future certification and validation suites

ampool altiscale pgemini CenturyLink DATATORRENT EMC²
Hortonworks IBM Infosys Linaro NEC Pivotal PLDT
sas splunk> squid syncsort Telstra TERADATA TOSHIBA
UNIFI vmware wanDISCO XIIIab Data Zettaset

IBM Open Platform foundational components

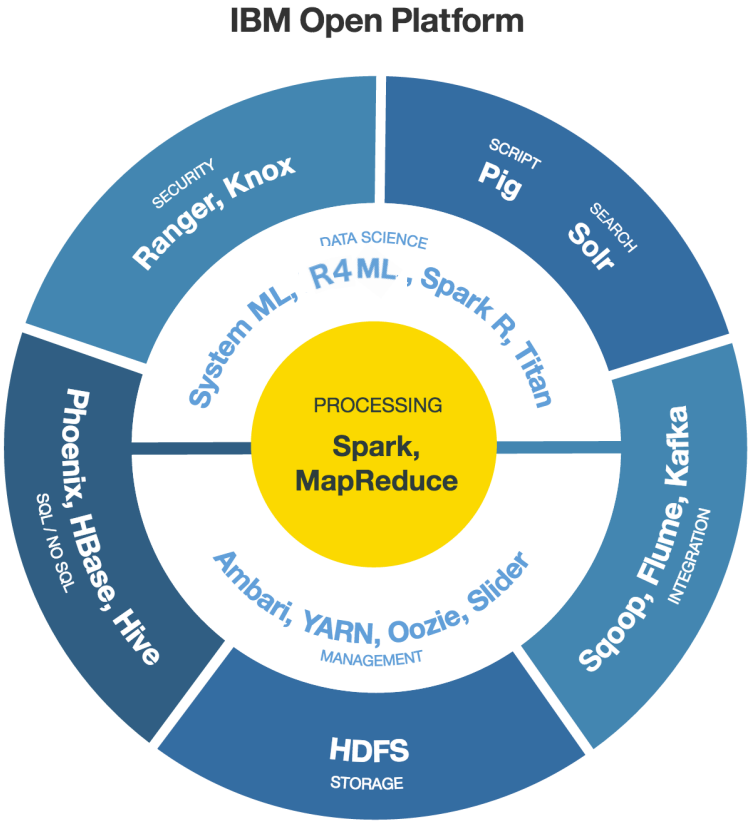
- **Apache Hadoop**
  - Distributed file system, popular API (**MapReduce**) for clustered computing
  - Originally designed for **batch** processing of massive data volumes, varied data formats
- **Apache Spark**
  - **General purpose**, high-speed data processing engine for clustered computing
  - **In-memory** processing, popular built-in libraries (e.g., machine learning)
  - **No built-in storage**. Attaches to other data stores (e.g., Hadoop Distributed File System)





"Better togetter"

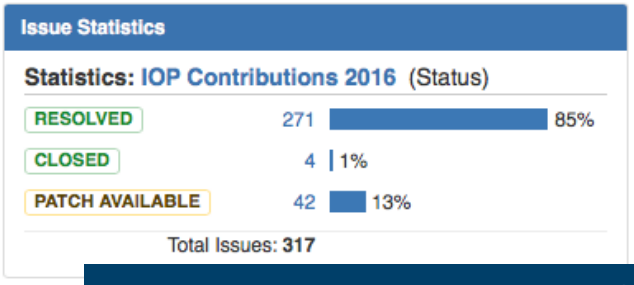
IBM Open Platform: a closer look

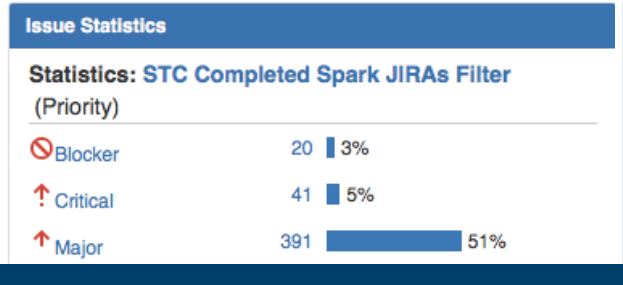


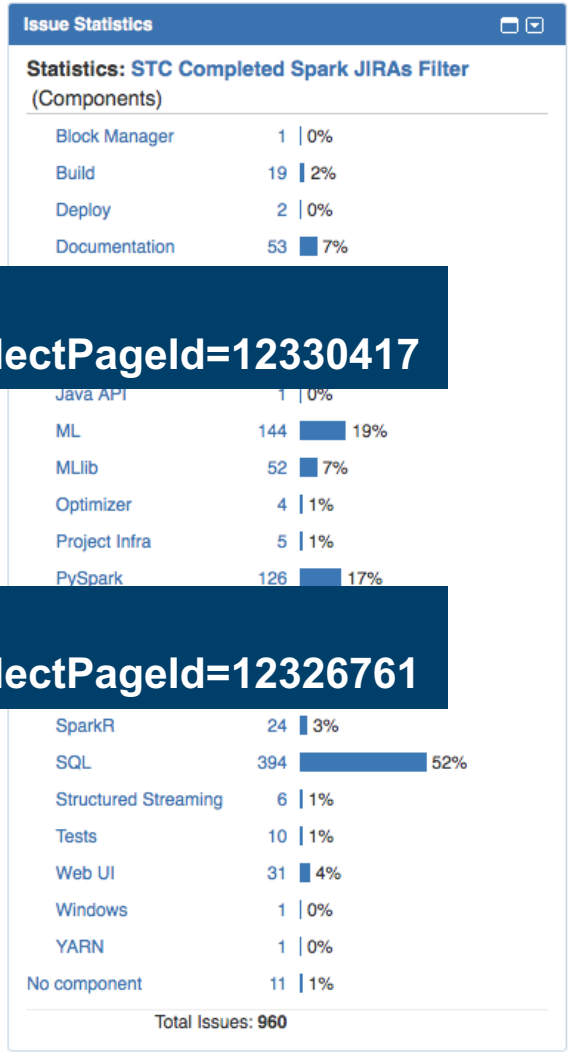
- **Compliant with ODPi runtime [OPEN]**
- **Timely updates as new open source versions released [CURRENT]**
- **Install only those components you want / need [MODULAR]**

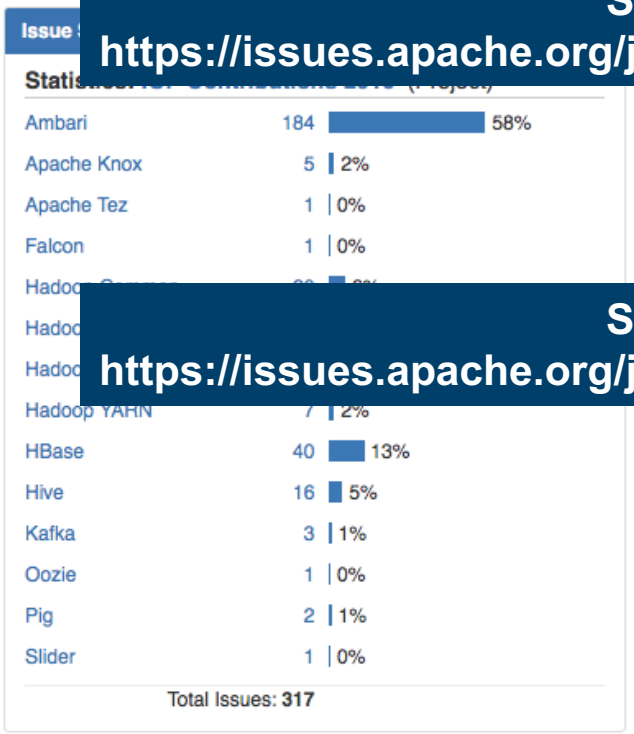
* Spark 2.1.0 available with new IOP 4.2.5 technical preview

IBM open source contributions









See it in action on YouTube:
https://issues.apache.org/jira/secure/Dashboard.jspa?selectPageId=12330417

See it in action on YouTube:
https://issues.apache.org/jira/secure/Dashboard.jspa?selectPageId=12326761

Snapshots taken Jan. 2017.
Latest content available online via Apache dashboards.

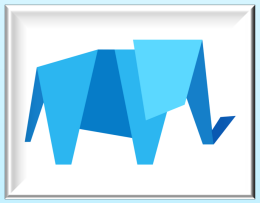
IOP relates to Hadoop; STC relates to Spark.

"Based on our experience, we found IBM Open Platform (IOP) ***surprisingly true to open source*** Apache Hadoop.

We support all major distributions of Hadoop including native Apache. Through our standard installation process we were able to have PepperData up and running with IOP ***in less than an hour***."

-- Alex Pierce
Sr. Field Engineer, PepperData

Overview of BigInsights



IBM-specific BigInsights features

Big SQL (industry standard SQL)
Text analytics
BigSheets (spreadsheet-style tool)

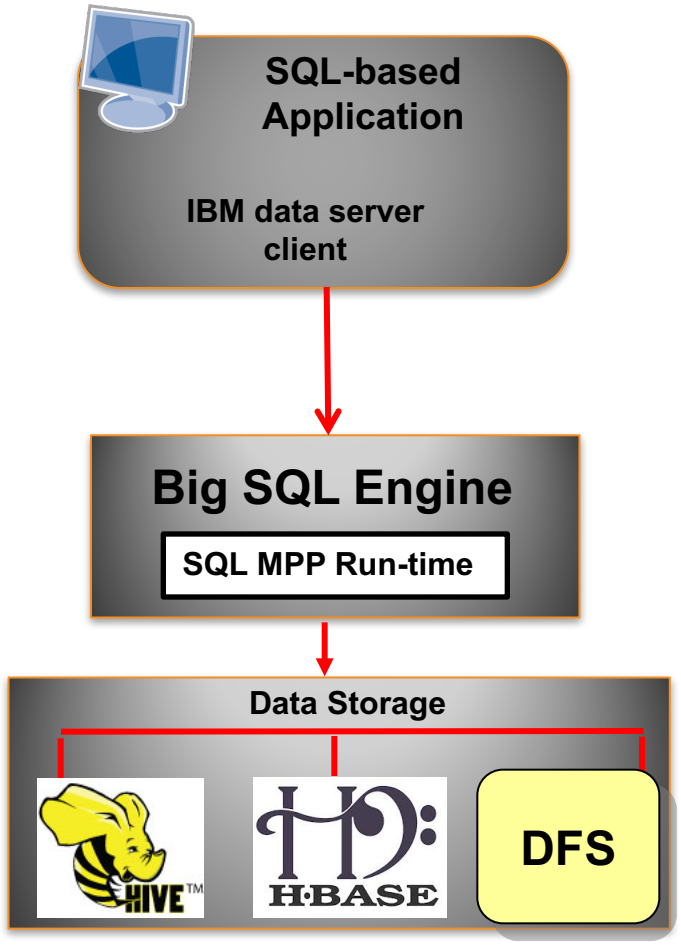
IBM Streams, Cognos (limited use licenses)

IBM Open Platform

100% open source platform compliant with ODPi
Apache Hadoop ecosystem
Apache Spark ecosystem

More value…

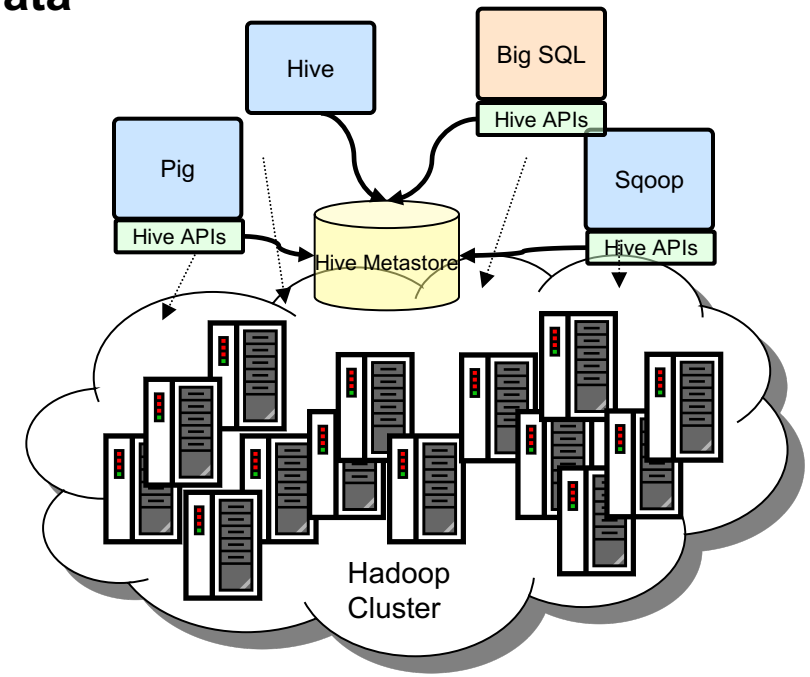
- **Optimization and performance**
  - IBM MPP engine **(C++)** replaces Java MapReduce layer
  - Continuous running daemons (no start up **latency**)
  - **Message** passing allow data to flow between nodes without persisting intermediate results
  - **In-memory** operations with ability to spill to disk (useful for aggregations, sorts that exceed available RAM)
  - Cost-based **query optimization** with140+rewrite rules (**Statistics**)
  - Resources are **automatically adjusted** based upon workload
    - **Self-tuning** memory manager that re-distributes resources across components dynamically.
- **Comprehensive, standard SQL (ANSI 2011)**
  - SELECT: joins, unions, aggregates, subqueries . . .
  - UPDATE/DELETE (HBase-managed tables)
  - GRANT/REVOKE, INSERT … INTO
  - SQL procedural logic (SQL PL)
  - Stored **procs, user-defined functions**
  - IBM data server **JDBC** and **ODBC** drivers (== **Informix** & **DB2**)
- **Security**
  - **roles**
  - **by row and/or column**
- **Loads data by command from:**
  - **Local or remote file system**
  - **RDBMSs (DB2, Netezza, Teradata, Oracle, MS-SQL, Informix) via JDBC connection**



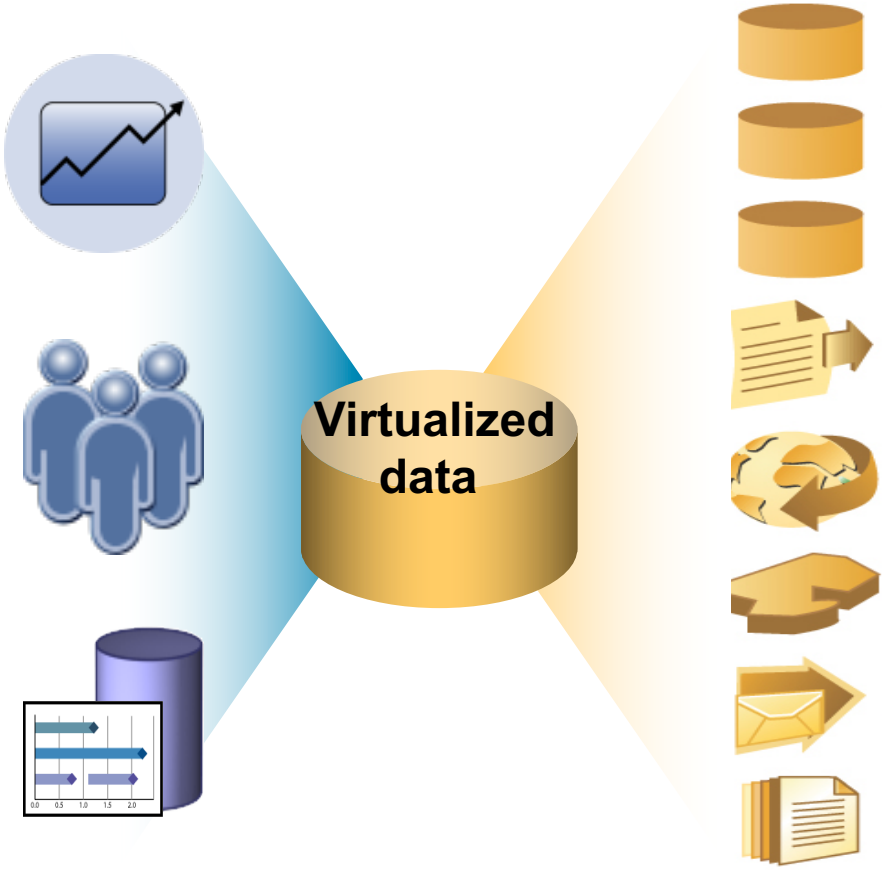
IBM Open Platform or Hortonworks Data Platform

IBM Big SQL Embraces Open Source HDFS file formats

- **Big SQL applies SQL to your existing Hadoop data**
  - **No propriety storage format**
  - Support Parquet, ORC, SEQ, delimited, Avro, HBase
- **A "table" is simply a view on your Hadoop data**
  - All data is Hadoop data
  - In files in HDFS
- **Table definitions shared with Hive**
  - The **Hive Metastore** catalogs table definitions
  - Reading/writing data logic is shared with Hive
  - Definitions can be shared across the Hadoop ecosystem
- **Data stored in Hive immediately query-able**



Big SQL query federation = virtualized data access



Transparent
- Appears to be one source
- Programmers don't need to know how / where data is stored

Heterogeneous
- Accesses data from diverse sources

High Function
- Full query support against all data
- Capabilities of sources as well

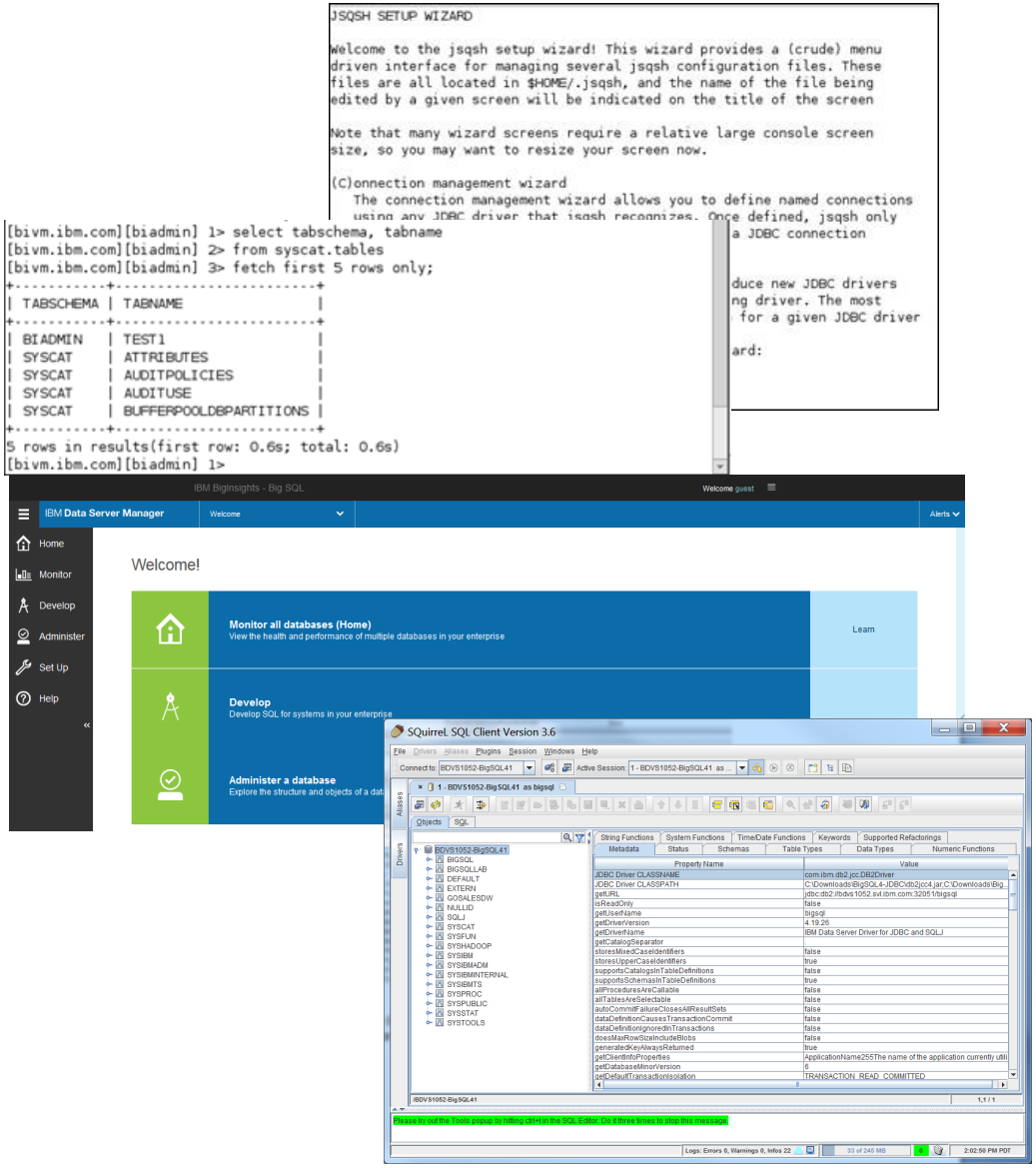
Autonomous
- Non-disruptive to data sources, existing applications, systems.

High Performance
- Optimization of distributed queries

INFORMIX DB2 TERADATA ODBC open database connectivity ORACLE NETEZZA Microsoft SQL Server 2014

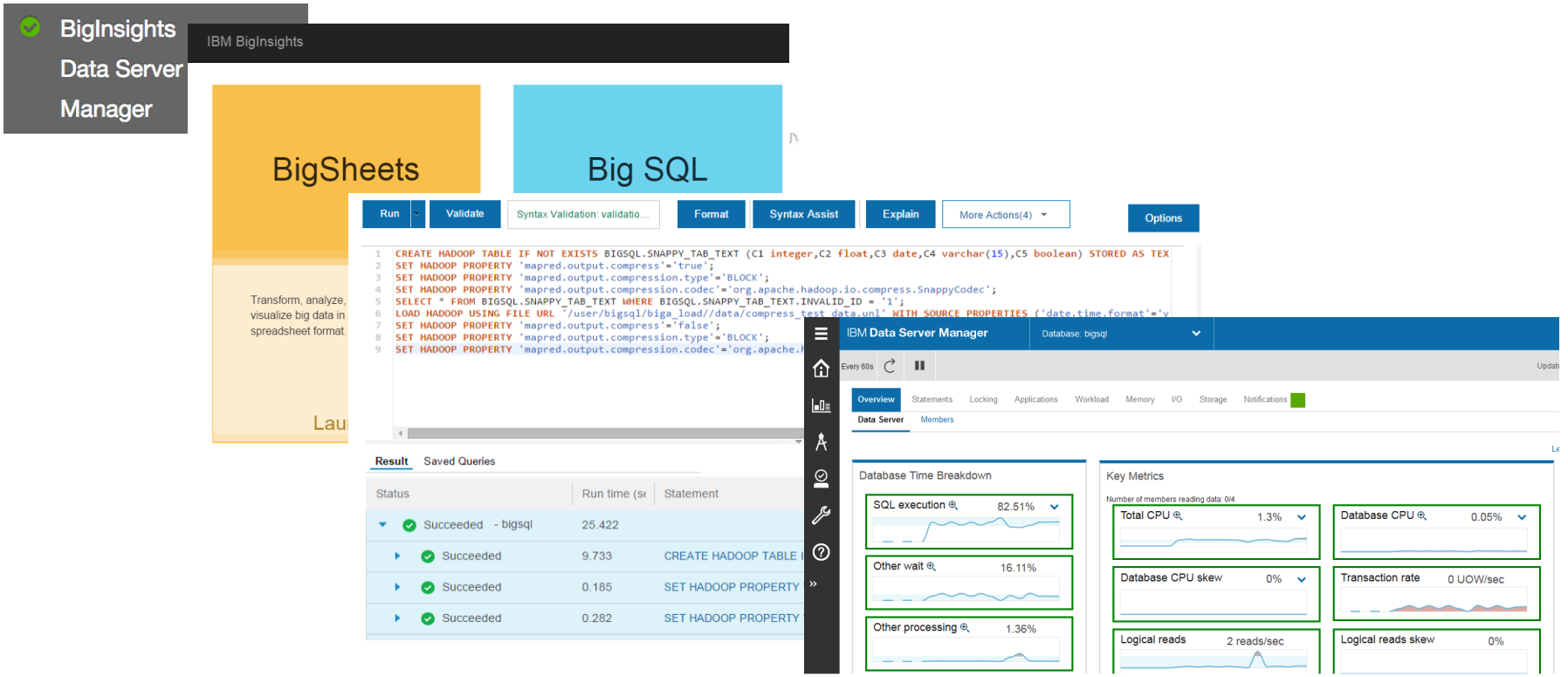
Invocation options

- **Tools that support IBM JDBC/ODBC driver**
- **Command-line interface: Java SQL Shell (JSqsh)**
- **Web tooling (Data Server Manager)**



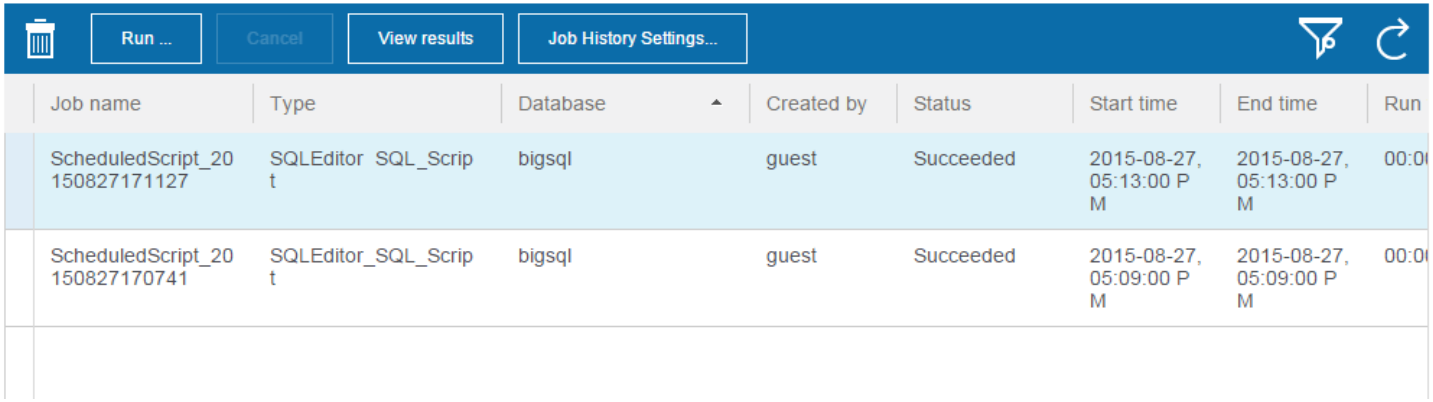
Data Server Manager

- Stand-alone service. Directly managed by Ambari
- Edit / execute Big SQL, monitor database, explore configuration, etc.
- Real time dashboards



Job Management

- Users can schedule a script or a task to be executed
- Job execution status and history view
- The detailed execution status results are available for display



About Big SQL performance: Hadoop-DS benchmark

- **TPC (Transaction Processing Performance Council)**
  - Formed August 1988
  - Widely recognized as most credible, vendor-independent SQL benchmarks
  - TPC-H and TPC-DS are the most relevant to SQL over Hadoop
    - R/W nature of workload not suitable for HDFS

- **Hadoop-DS benchmark: BigInsights, Hive and Cloudera**
  - Original benchmark **run by IBM & reviewed by TPC certified auditor** in Oct 2014
  - Updates (including results presented here) not been reviewed by TPC certified auditor
  - Based on TPC-DS. Key deviations
    - No data maintenance or persistence phases (not supported across all vendors)
  - Common set of queries across all solutions
    - Subset that ***all*** vendors can successfully execute at scale factor
    - Queries are not cherry picked
  - Most complete TPC-DS like benchmark executed so far
  - *Analogous to porting a relational workload to SQL on Hadoop*

Big SQL runs all queries without extensive modifications

Other environments require significant effort at scale

BigSQL 4.1

Query 1	Query 34	Query 67
Query 2	Query 35	Query 68
Query 3	Query 36	Query 69
Query 4	Query 37	Query 70
Query 5	Query 38	Query 71
Query 6	Query 39	Query 72
Query 7	Query 40	Query 73
Query 8	Query 41	Query 74
Query 9	Query 42	Query 75
Query 10	Query 43	Query 76
Query 11	Query 44	Query 77
Query 12	Query 45	Query 78
Query 13	Query 46	Query 79
Query 14	Query 47	Query 80
Query 15	Query 48	Query 81
Query 16	Query 49	Query 82
Query 17	Query 50	Query 83
Query 18	Query 51	Query 84
Query 19	Query 52	Query 85
Query 20	Query 53	Query 86
Query 21	Query 54	Query 87
Query 22	Query 55	Query 88
Query 23	Query 56	Query 89
Query 24	Query 57	Query 90
Query 25	Query 58	Query 91
Query 26	Query 59	Query 92
Query 27	Query 60	Query 93
Query 28	Query 61	Query 94
Query 29	Query 62	Query 95
Query 30	Query 63	Query 96
Query 31	Query 64	Query 97
Query 32	Query 65	Query 98
Query 33	Query 66	Query 99

Impala 2.2

Query 1	Query 34	Query 67
Query 2	Query 35	Query 68
Query 3	Query 36	Query 69
Query 4	Query 37	Query 70
Query 5	Query 38	Query 71
Query 6	Query 39	Query 72
Query 7	Query 40	Query 73
Query 8	Query 41	Query 74
Query 9	Query 42	Query 75
Query 10	Query 43	Query 76
Query 11	Query 44	Query 77
Query 12	Query 45	Query 78
Query 13	Query 46	Query 79
Query 14	Query 47	Query 80
Query 15	Query 48	Query 81
Query 16	Query 49	Query 82
Query 17	Query 50	Query 83
Query 18	Query 51	Query 84
Query 19	Query 52	Query 85
Query 20	Query 53	Query 86
Query 21	Query 54	Query 87
Query 22	Query 55	Query 88
Query 23	Query 56	Query 89
Query 24	Query 57	Query 90
Query 25	Query 58	Query 91
Query 26	Query 59	Query 92
Query 27	Query 60	Query 93
Query 28	Query 61	Query 94
Query 29	Query 62	Query 95
Query 30	Query 63	Query 96
Query 31	Query 64	Query 97
Query 32	Query 65	Query 98
Query 33	Query 66	Query 99

Key points

- With Impala some queries needed to be re-written, some significantly
- Re-writing queries in a benchmark scenario where results are known is one thing – doing this against real databases in production is another

Works without modification

Minor modification

Extensive modification

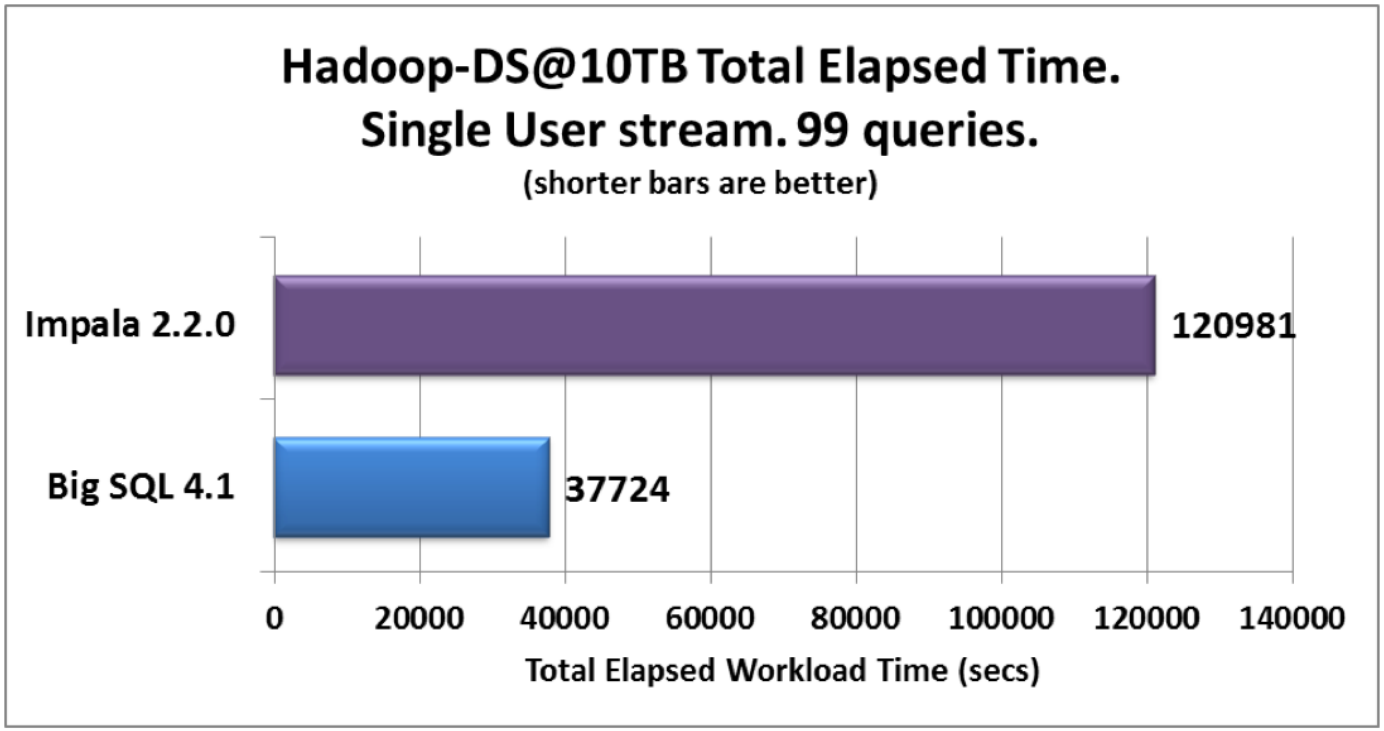
Not working

 Results for 10TB scale shown here

Hadoop-DS benchmark single user performance 10TB

Big SQL is 3.2x faster than Impala

for single query stream using 99 common queries

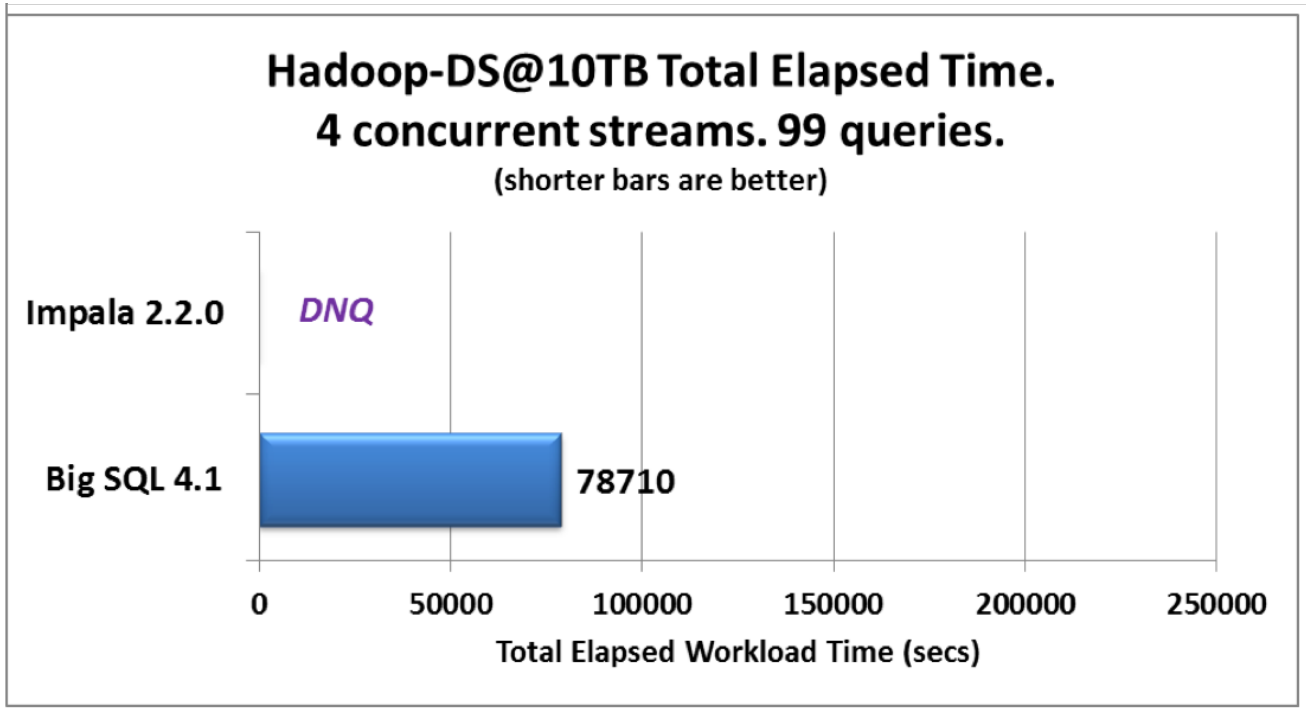


Based on IBM internal tests comparing BigInsights Big SQL and Cloudera Impala running on **identical hardware**. The test workload was based on the latest revision of the TPC-DS benchmark specification at 10TB data size. Successful executions measure the ability to execute queries a) directly from the specification without modification, b) after simple modifications, c) after extensive query rewrites. All minor modifications are either permitted by the TPC-DS benchmark specification or are of a similar nature. Development effort measured time required by a skilled SQL developer familiar with each system to modify queries so they will execute correctly. Performance test measured total workload elapsed time of single and 4 concurrent users executing all 99 queries across all 3 systems at 10TB data size. Results may not be typical and will vary based on actual workload, configuration, applications, queries and other variables in a production environment.

Hadoop-DS benchmark multi user performance 10TB

With 4-streams Big SQL..

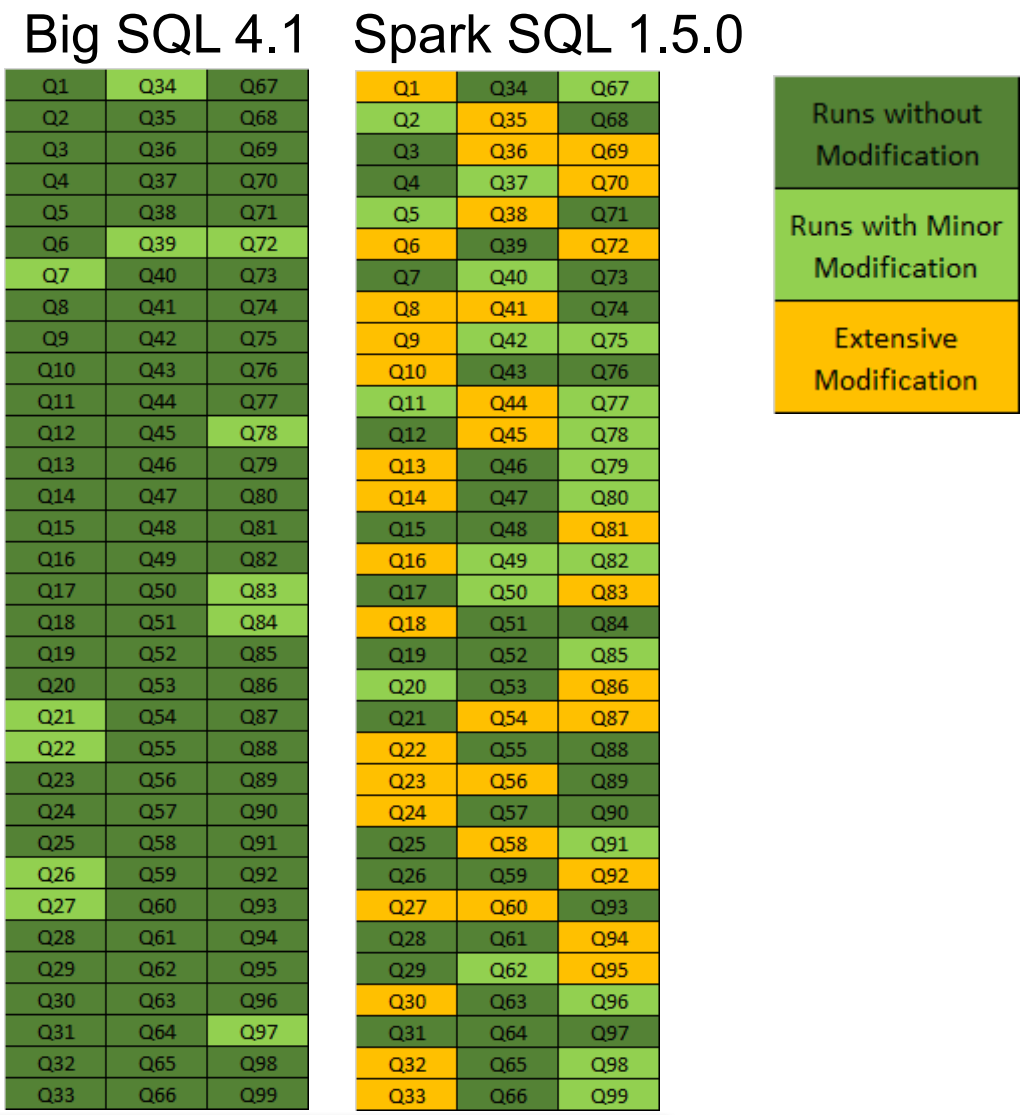
For 4 query streams using 99 common queries



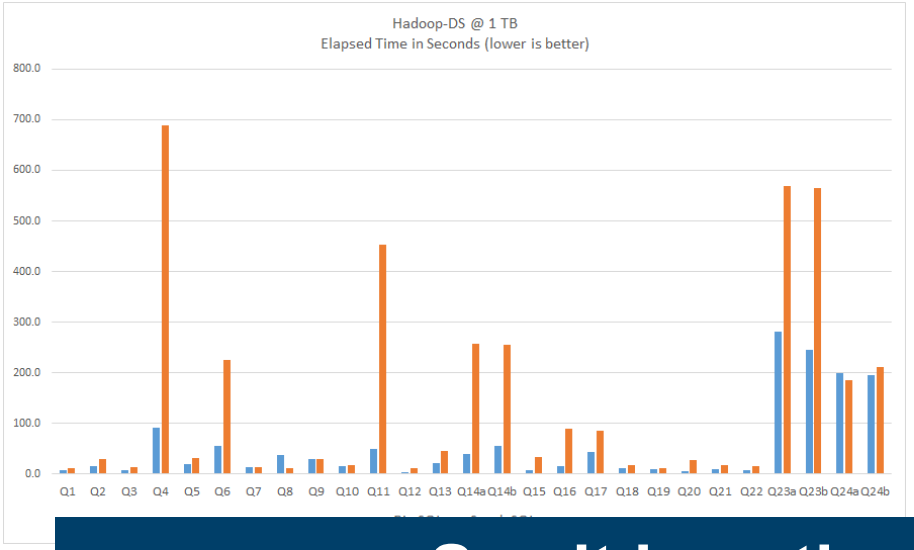
Based on IBM internal tests comparing BigInsights Big SQL and Cloudera Impala running on **identical hardware**. The test workload was based on the latest revision of the TPC-DS benchmark specification at 10TB data size. Successful executions measure the ability to execute queries a) directly from the specification without modification, b) after simple modifications, c) after extensive query rewrites. All minor modifications are either permitted by the TPC-DS benchmark specification or are of a similar nature. Development effort measured time required by a skilled SQL developer familiar with each system to modify queries so they will execute correctly. Performance test measured total workload elapsed time of single and 4 concurrent users executing all 99 queries across all 3 systems at 10TB data size. Results may not be typical and will vary based on actual workload, configuration, applications, queries and other variables in a production environment.

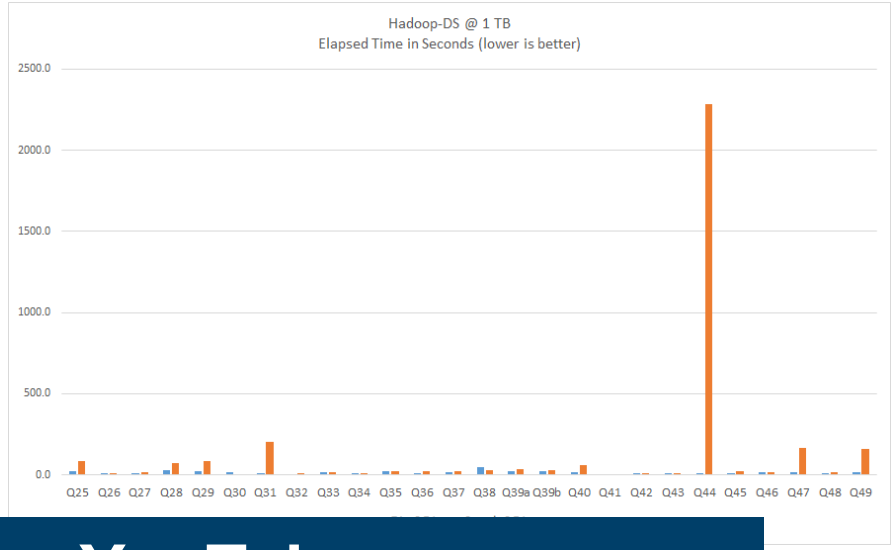
Big SQL runs more SQL out-of-box

Big SQL is the ***only*** engine that can execute all 99 queries with minimal porting effort

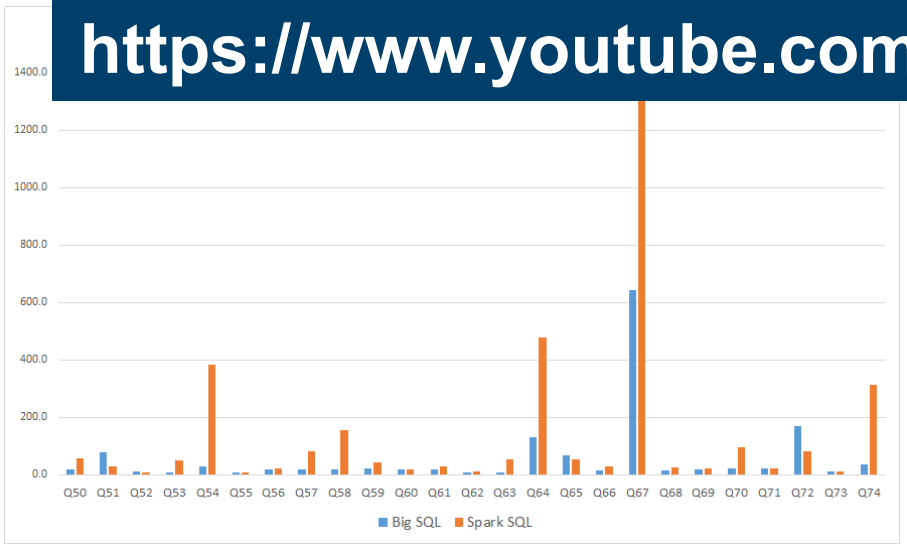


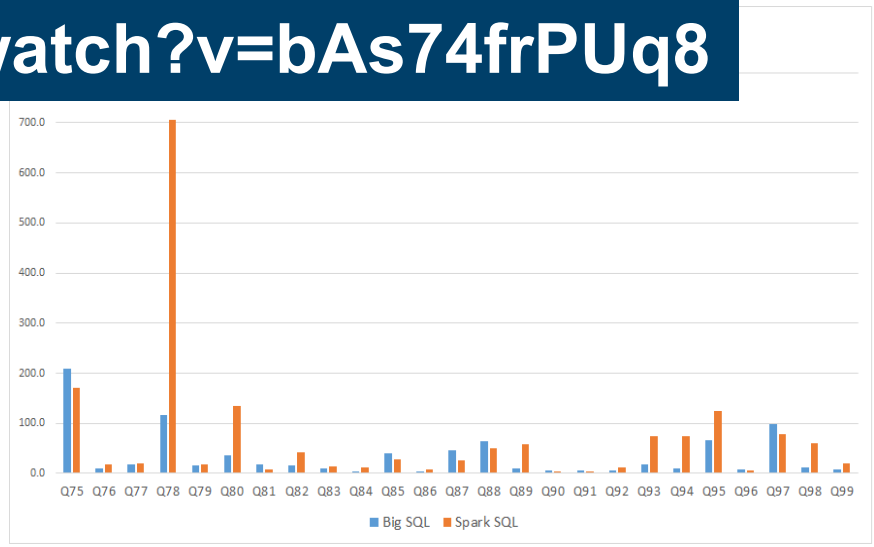
Big SQL 4.1 vs. Spark 1.5.0, Single Stream @ 1TB



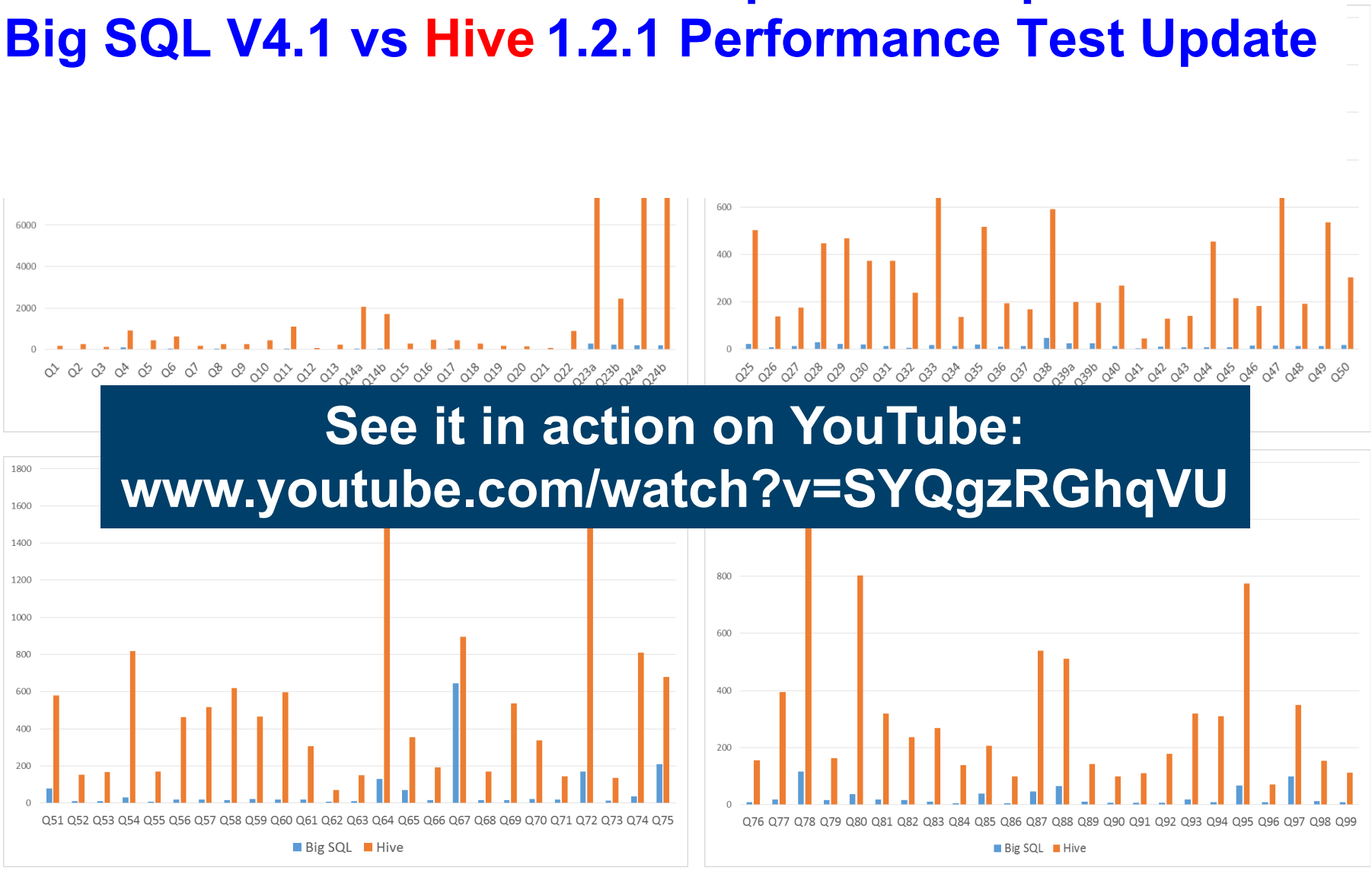


See it in action on YouTube:
https://www.youtube.com/watch?v=bAs74frPUq8





Announced at Strata + Hadoop World Sept 2015: Big SQL V4.1 vs Hive 1.2.1 Performance Test Update



See it in action on YouTube: www.youtube.com/watch?v=SYQgzRGhqVU

Distinguishing characteristics

Application Portability & Integration

Data shared with Hadoop ecosystem

Comprehensive file format support

Superior enablement of IBM and Third Party software

Performance

Modern MPP runtime

Powerful SQL query rewriter

Cost based optimizer

Optimized for concurrent user throughput

Results not constrained by memory

Rich SQL

Comprehensive SQL Support

IBM SQL PL compatibility

Extensive Analytic Functions

Federation

Distributed requests to multiple data sources within a single SQL statement

Main data sources supported:
DB2, Teradata, Oracle, Netezza, Informix, SQL Server, ODBC

Enterprise Features

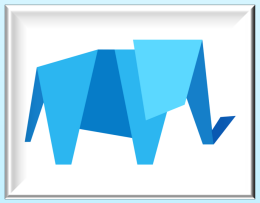
Advanced security/auditing

Resource and workload management

Self tuning memory management

Comprehensive monitoring

Overview of BigInsights



IBM-specific BigInsights features

Big SQL (industry standard SQL)
Text analytics
BigSheets (spreadsheet-style tool)

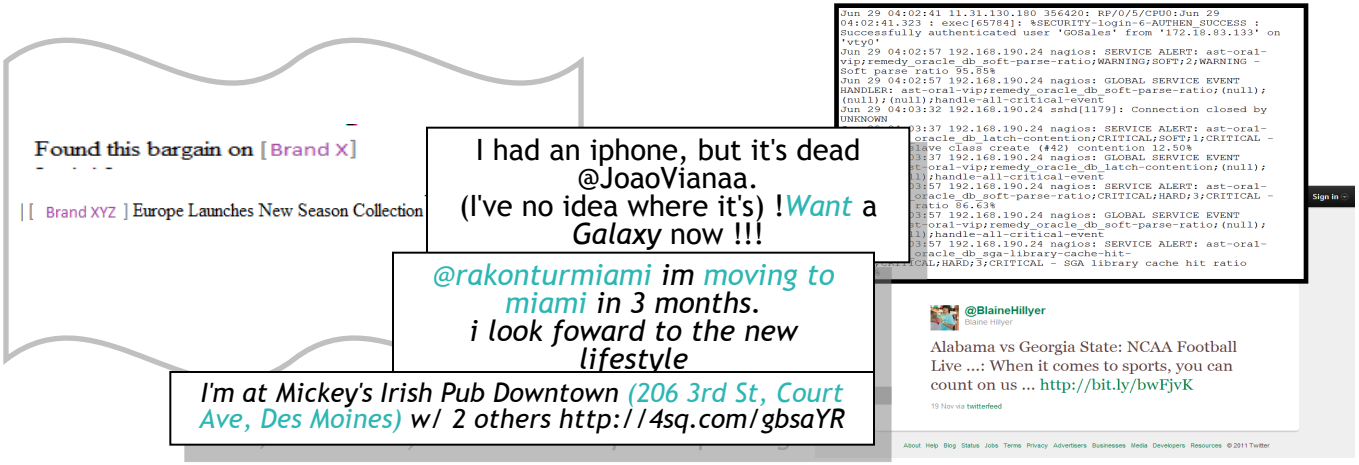
IBM Streams, Cognos (limited use licenses)

IBM Open Platform

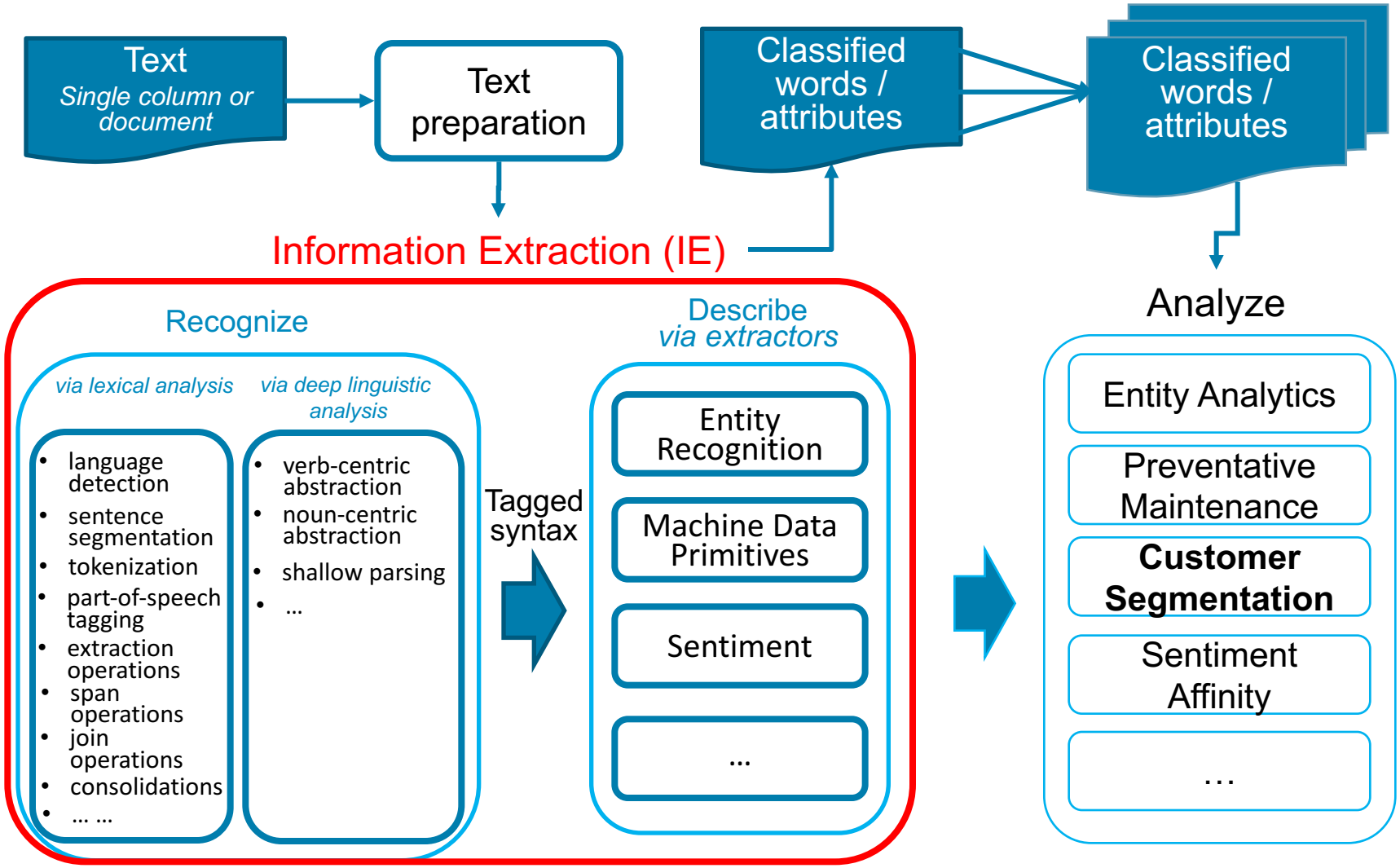
100% open source platform compliant with ODPi
Apache Hadoop ecosystem
Apache Spark ecosystem

Text analytics

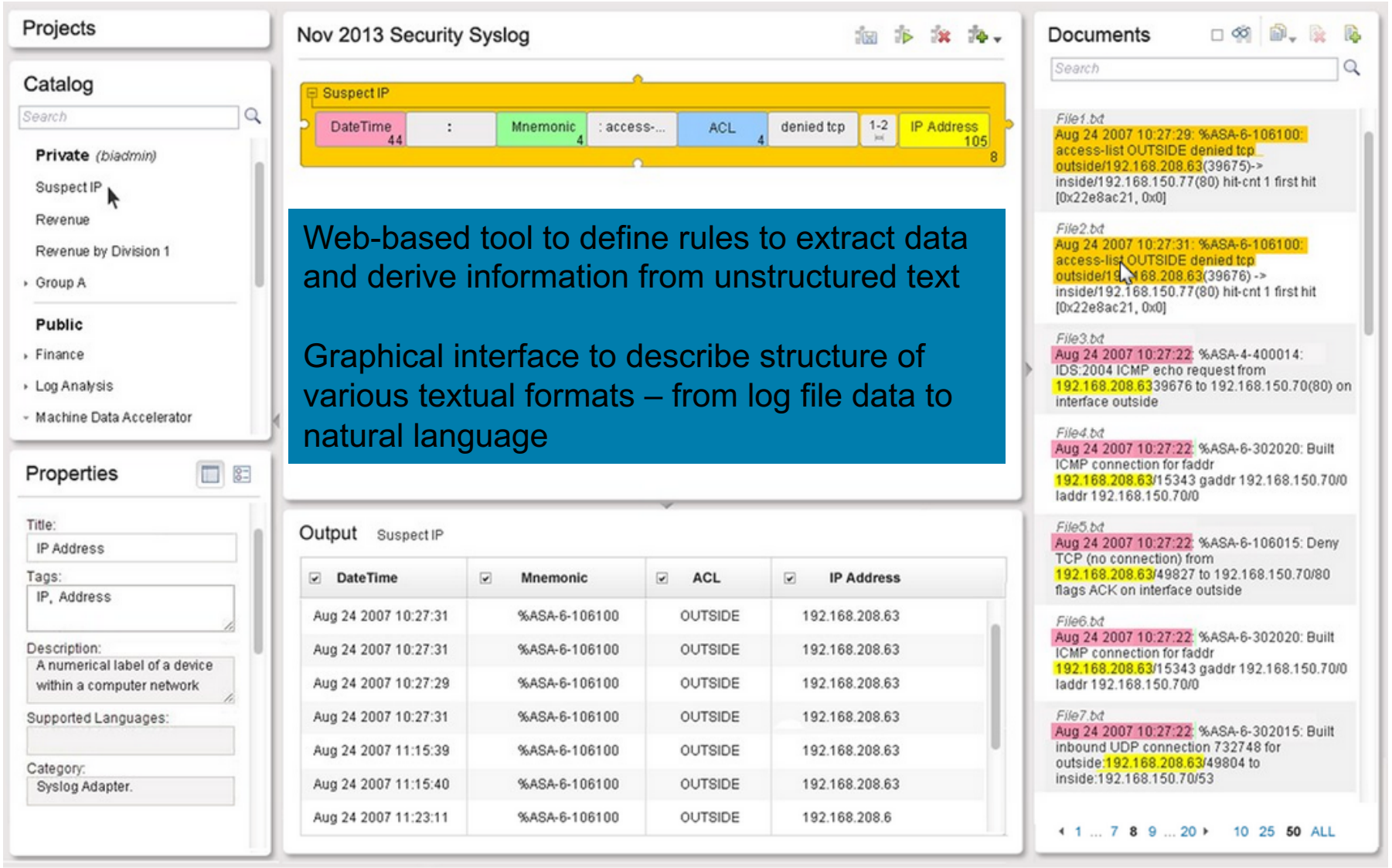
- **Distills structured info from unstructured text**
  - Sentiment analysis
  - Consumer behavior
  - Illegal or suspicious activities
  - …
- **Parses text and detects meaning with annotators**
- **Understands the context in which the text is analyzed (Required)**
- **Features pre-built extractors for names, addresses, phone numbers, etc.**



Extracting information from text

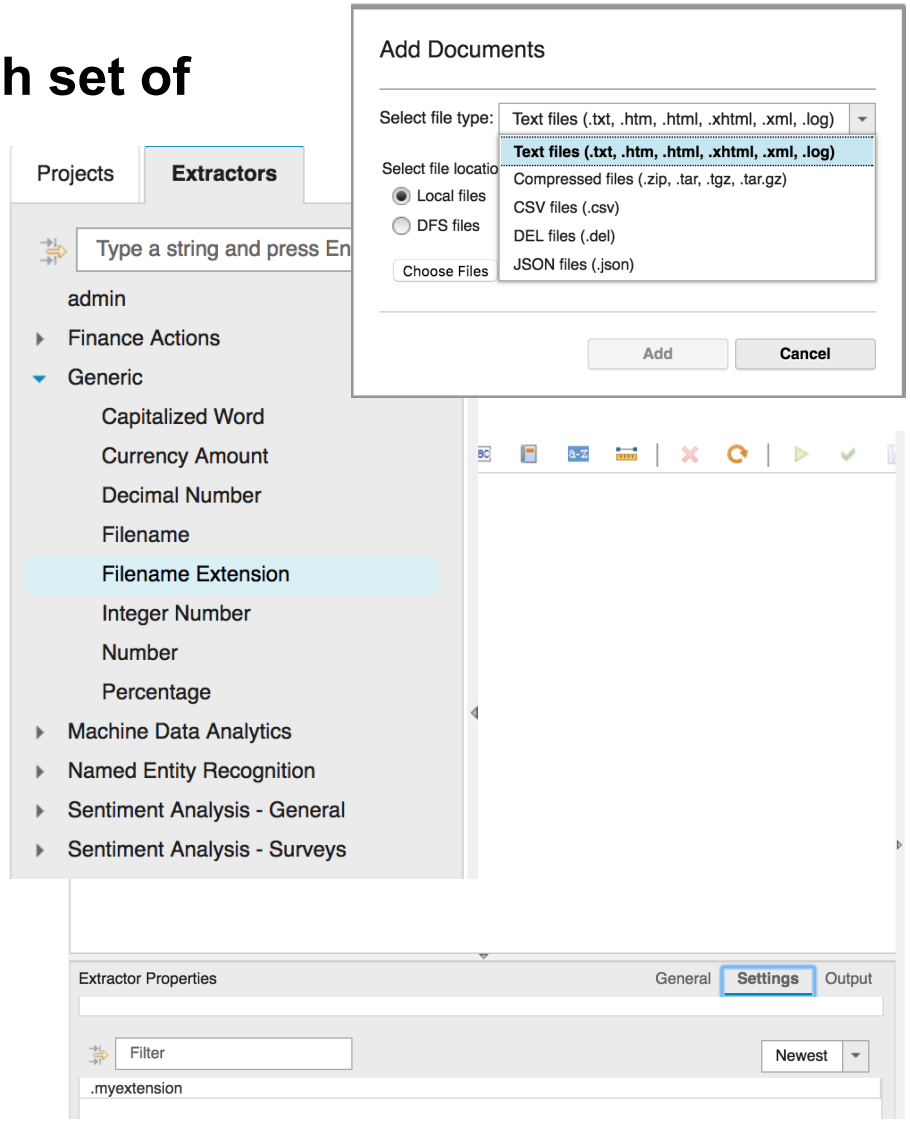


Text analytics tooling



Input >> Pre-built text extractors >> Output

- **The extractor library contains a rich set of pre-built extractors**
  - Finance actions
  - Named Entities
  - Generic
  - Machine Data
  - Sentiment Analysis
- **You can control output properties**
  - Output columns and names
  - Row filters
- **Some pre-built extractors can be customized**
  - Add / remove dictionary terms



IBM BigInsights – Text Analytics

Extract information from unstructured sources for business insight

MrSurvey: Hello! Where are you from?

Customer: I'm from [Alicante]

MrSurvey: Do you think that the public transport will be improve?

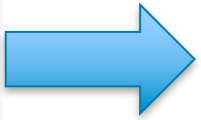
Customer: Well, [yes], the [frequency] could be improved

MrSurvey: And do you think the routes should be expanded?

Customer: [Yes], Most of the routes are focus to north zone, eg: It could be a good way reach [Elche]

Agent: How much you pay for a ticket between two cities 30km far away (eg: Alicante to Novelda)?

Customer: It could be [10] euros.

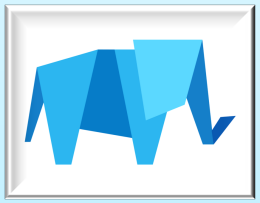


LOCATION	IMPROVEMENT	WHAT	EXPAND	EXPAND_CITY	30K_DESIRED_PRICE
Alicante	1	frequence	1	Elche	10
Elche	0	route	1	Alicante	15
Torrevieja	1	price	0		12

BigInsights Text Analytics

- **High Performance rule based Information Extraction Engine**
  - Developed at IBM Research since 2004
- **Highly scalable solution available for at-rest and in-motion analytics**
  - Deploy Extractors to BigInsights and Streams
- **Pre-built extractors, and toolkit to build custom Extractors**
  - Rich Extractor library supports multiple languages
  - Declarative Information Extraction (IE) system based on an algebraic framework and AQL language
- **Tooling to help build, test, and refine Extractors**
  - Simple drag and drop web UI for visual assembly by Data Scientists
- **IBM's strategic Information Extraction Engine**
  - Embedded in several IBM products

Overview of BigInsights



IBM-specific BigInsights features

Big SQL (industry standard SQL)
Text analytics
BigSheets (spreadsheet-style tool)

IBM Streams, Cognos (limited use licenses)

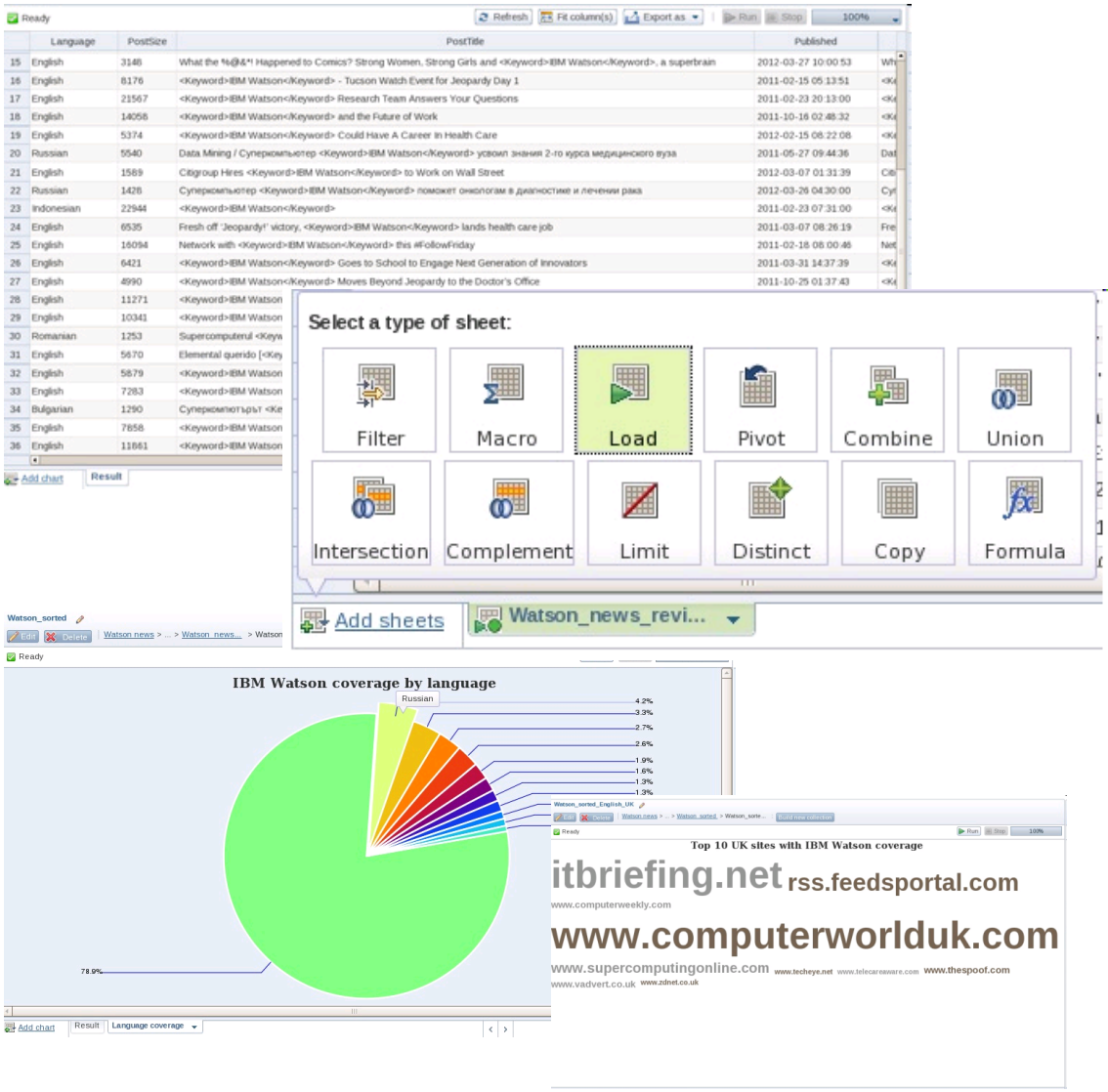
IBM Open Platform
100% open source platform compliant with ODPi
Apache Hadoop ecosystem
Apache Spark ecosystem

Spreadsheet-style analysis (BigSheets)

- **Web-based analysis and visualization**

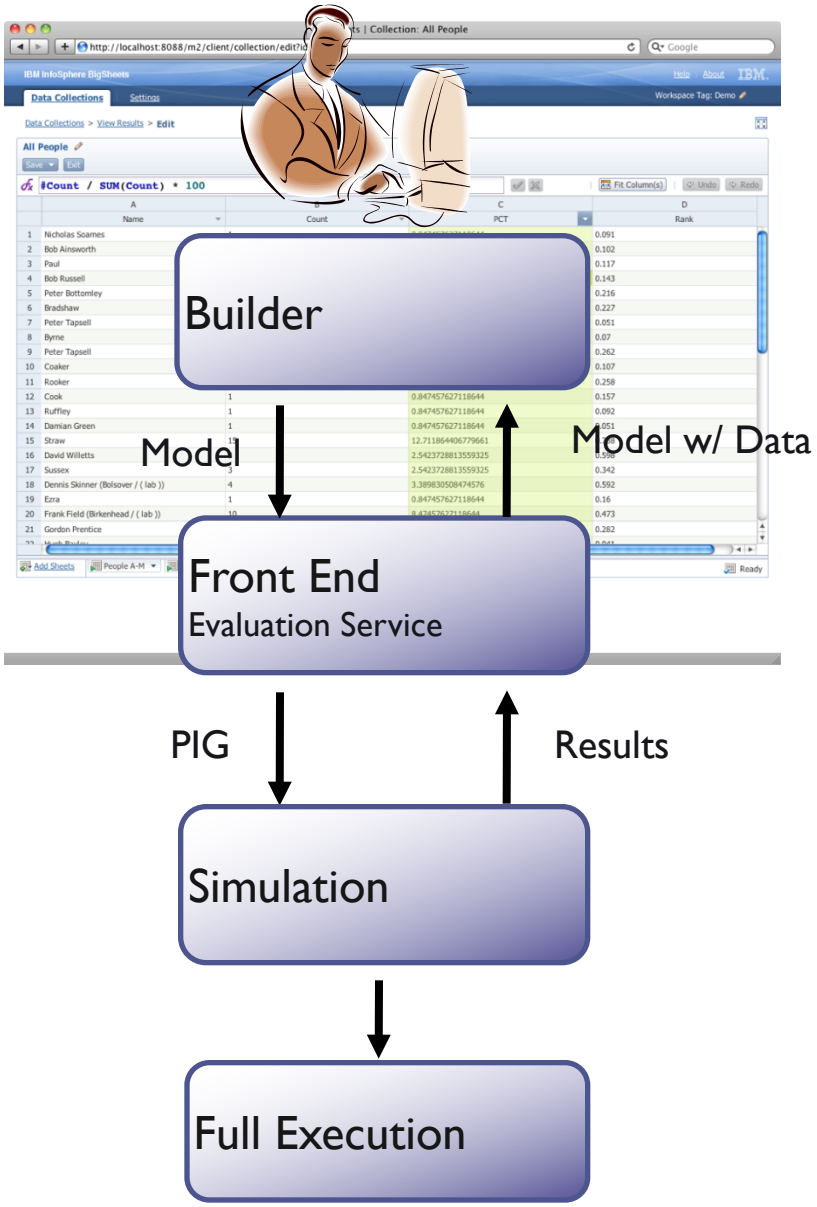
- **Spreadsheet-like interface**
  - Explore, manipulate data without writing code
  - Invoke pre-built functions (174, ABS, LOWER, …)
  - Generate charts
  - Export results of analysis
  - Create custom plug-ins
  - …

- **REST API**
  - Create master workbook
  - Run workbook
  - …

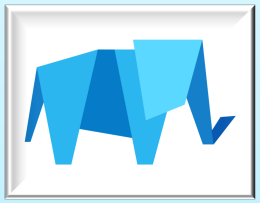


Working with BigSheets

- **Create workbook for data**
- **Customize workbook through graphical editor and built-in functions**
  - Filter data
  - Apply functions / macros / formulas
  - Combine data from multiple workbooks
- **"Run" workbook: apply work to full data set**
- **Explore results in spreadsheet format and/or create charts**
- **Optionally, export your data**



Overview of BigInsights



IBM-specific BigInsights features

Big SQL (industry standard SQL)
Text analytics
BigSheets (spreadsheet-style tool)

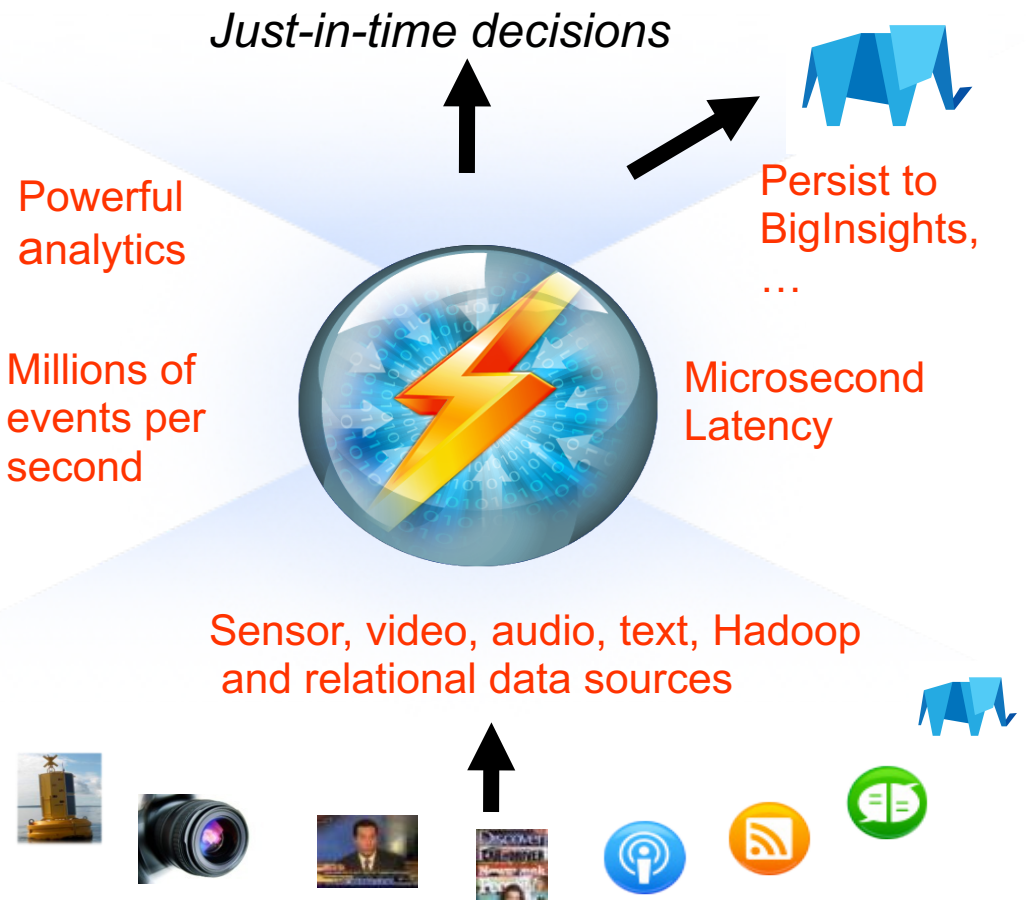
IBM Streams, Cognos (limited use licenses)

IBM Open Platform

100% open source platform compliant with ODPi
Apache Hadoop ecosystem
Apache Spark ecosystem

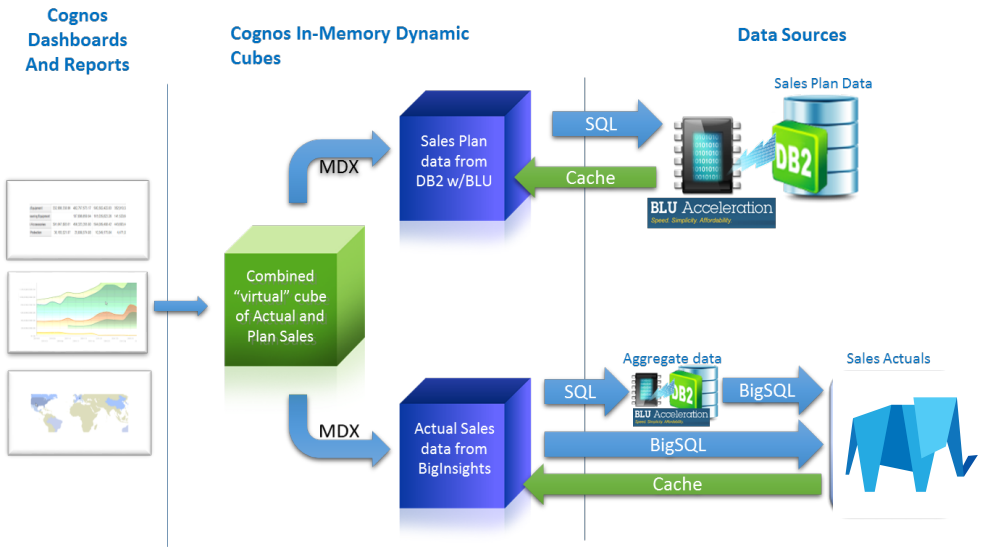
Limited use license: IBM Streams

- Platform for real-time Big Data analytics
  - "Data in motion"
  - Gigabytes+ per second or more
  - Terabyte+ per day
  - All kinds of data
  - Insights in microseconds
- Connectivity to varied data sources



Limited use license: Cognos BI

- Model, explore, analyze data from many sources
- Visualize and report on results
- **Connection to BigInsights via Big SQL**
- In-memory dynamic views cache data in Cognos for quick data access
- Part of IBM BigInsights for Apache Hadoop



COGNOS
Better Decisions Every Day™

Demo: https://www.youtube.com/watch?v=yxnoGrK6PSY

Thinking cloud? Think IBM!

FASTER INNOVATION

Try-Before-You-Buy

BETTER ECONOMICS

Buy only what you need.
Start small and grow.

LOWER RISK OF FAILURE

\+ Lower Skill
Less Cost

EQUALS

REDUCING RISK!

BigInsights on Cloud offering options

Enterprise

- **Dedicated** hardware
- Small/Med/Large dedicated hosts
- **Monthly** subscription
- IBM Open Platform
- IBM BigInsights
- VPN option
- Secure: ISO27K1, SOC2, HIPAA certified

Basic

- **Shared** hardware
- Virtual nodes: 4 core, 24GB, 244GB disk
- Object Storage service integration
- **Hourly** pay-as-you-go
- Free usage for trial period
- IBM Open Platform only

IBM BigInsights on cloud

Build

- **Ready-to-run** Hadoop clusters in the cloud
- IBM Open Platform - 100% open source Hadoop; will align with ODP
- Based on proven, performant reference architectures

Manage

- Key platform components monitored for availability
- Hadoop, OS and BigInsights **patched and maintained**
- Ambari cluster manager for complete control

Support

- **24x7** cloud operations and support team
- Access to deep Hadoop expertise
- Faster time to problem resolution

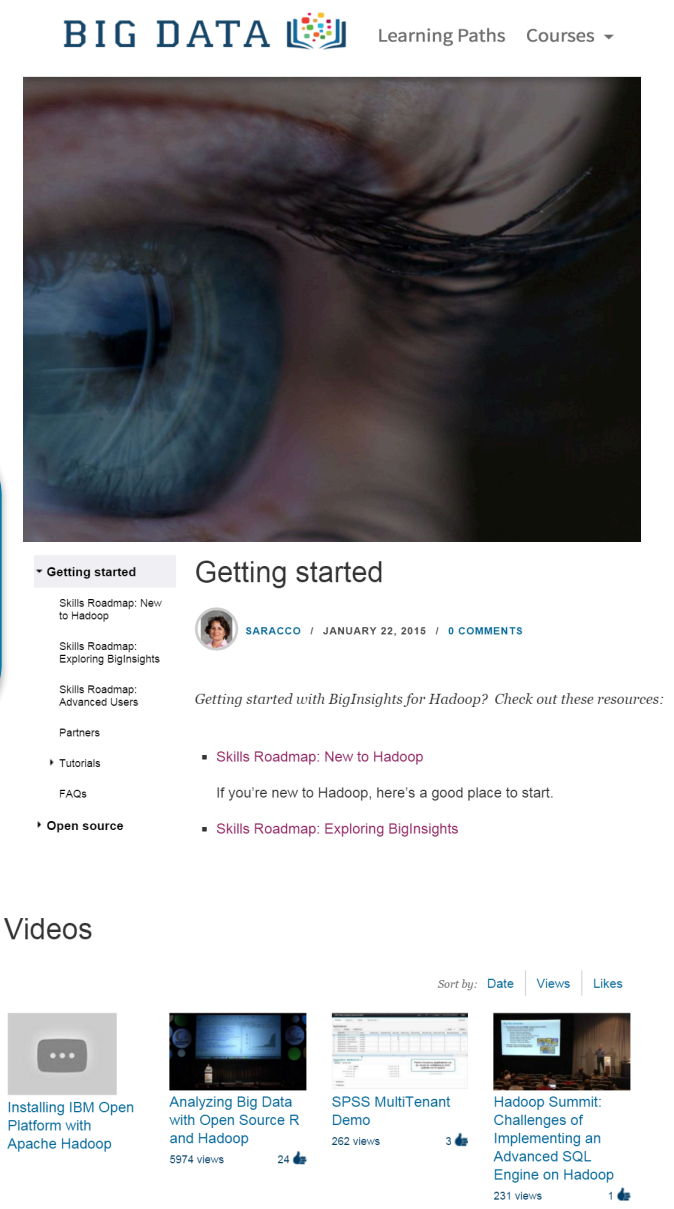
Protect

- Deployed in world-class, secure **SoftLayer (Enterprise)** data centers
- Dedicated physical machines
- **Certified** SSAE SOC2 Type 2, SOC3, ISO 27001

http://www.ibm.com/cloud
http://www.bluemix.net

Want to learn more?

- **Follow tutorials, videos, and more**
  - **WOW!:** BigDataUniversity
    - https://bigdatauniversity.com/
- **Links all available from HadoopDev**
  - https://developer.ibm.com/hadoop/
- **Download Quick Start image**
  - **or try on Bluemix Cloud**




IBM big data • IBM big data • IBM big data
THINK
BIG
IBM big data • IBM big data • IBM big data
IBM big data • IBM big data
IBM big data • IBM big data