Apache Spark 2.0 GA
The General Engine for Modern Analytic Use Cases
Apache Spark Drives Business Innovation

Apache Spark is driving new business value that is being harnessed by technology forward organizations.

Driving Customer Insights
- Next Best Offer (Machine Learning)
- Churn Analysis
- Click-Stream (Stream Processing)

Improving Products and Service Efficiencies
- Streaming from IOT Sources
- Connected Products/Services Analysis
- Proactive/Predictive Maintenance

Lower Business Risks
- Risk Modeling & Analysis
- Network Threat Detection
Spark Addresses Common Limitations

Access and Usability
One of the key advantages of Apache Spark is the intuitive and flexible API for big-data processing, available in popular programming languages. Prior to Apache Spark, users had access to very limited in-flexible abstractions for processing large distributed data, with poor support outside java.

Data Processing Performance
Mapreduce made big strides in enabling cost effective batch processing of large volumes of data. However, businesses continue to see a need to shorten data processing windows and consume data faster, requiring a new framework with significantly better performance.

Machine Learning at Scale
Data Science and Machine Learning on big-data are exciting areas of focus. However that requires libraries and that enable building models on large distributed data and APIs that allow flexible exploration of data.
Apache Spark
Fast and flexible general purpose data processing for Hadoop

Data Engineering
Stream Processing
Data Science & Machine Learning

Unified API and processing Engine for large scale data
Spark Use Cases

Top use cases for Spark

- DATA PROCESSING (ETL) (55%)
- REAL-TIME PROCESSING (44%)
- DATA SCIENCE (33%)
- MACHINE LEARNING (33%)

71% are employing Spark for Data Science

Top Use Cases Data Processing (55%), Real-Time Stream Processing (44%), Exploratory Data Science (33%) and Machine Learning (33%).

3 out of 8 are employing Spark in data science research
Why Spark at Cloudera?
The Most Apache Spark Experience

Cloudera is the “stress free” choice for Spark
• Support: Proactive Support for Spark workloads
• Expertise: Most Spark users trained. Robust development community.
• Experience: First to ship and support. Most customers running Spark of any commercial Hadoop Distribution.

Cloudera lives where your data lives
• Run Spark On-prem or in the Public Cloud

Cloudera makes Spark enterprise hardened
• Comprehensive Management and Alerting
• End to End Security and Governance
• Better Multi-tenancy operation for multiple workloads

Out-of-the-box ready for end to end use cases
• Spark with supported seamless integrations with other big-data tools (Kafka, Hbase, Kudu, etc)
57% have adopted Cloudera Spark for their most important use case, vs. 26% Hortonworks, 22% an Apache download, and 7% Databricks.

48% of respondents said they most commonly use Spark with HBase and 41% of respondents said they use Spark with Kafka.

**Source: Tejena Group Apache Spark Market Survey 2016** http://tanejagroup.com/profiles-reports/request/apache-spark-market-survey-cloudera-sponsored-research#.WCCdPC0rK70
The One Platform Initiative

Management
Leverage Hadoop-native resource management

Security
Full support for Hadoop security and beyond

Scale
Spark at Petabyte scale

Streaming
Performance, simplification & easy-management of streaming workloads

Cloud
Elastic transient workloads
Three Core Enterprise Applications

Data Engineering & Science
Process data, develop & serve predictive models

Analytic Database
ELT, reporting, exploratory business intelligence

Operational Database
Build data-driven applications to deliver real-time insights

OPERATIONS
PROCESS, ANALYZE, SERVE
UNIFIED SERVICES
STORE
INTEGRATE

DATA MANAGEMENT
Cloudera’s Data Engineering Solution

Navigator
- Audit, lineage, encryption, key management, & policy lifecycles

Search
- Interactive search and immediate exploration

Cloud Deployment
- Easy deployment and flexible scaling

Hive-on-Spark
- Large-scale ETL & batch processing engine

Spark
- Modern Real-time Analytics Engine

Operating System, Analyze, Serve
- Unified Services
- Store
- Integrate

Data Management

Multi-Storage, Multi-Environment
Data Processing
Common Limitations

Poor Cloud Design
ETL and Batch Processing workloads need to utilize large amounts of compute but for only a window of time. This causes organizations to over provision to meet demands of the job while the environment lays dormant a majority of the time producing poor ROI.

Poor Performance
ETL and data processing takes too long and often excludes important data sources that are needed to extract real value from data collected. Traditional platforms only leverage structured data but increasingly the data needed to offer true intelligence varies in format and delivery.

Limited Data Formats
Traditional platforms only leverage structured data and require a strategic approach to schema design. Introducing new data (unstructured, time series, nested, log data) is often complex if not impossible. This causes analysis to be limited to only data extracted from core systems.
Data Processing with Spark
Process large scale unstructured and structured data in the same application

Powerful and flexible higher order functions for arbitrary processing of structured or unstructured data

- map
- flatMap
- filter
- union
- reduceByKey
- groupBy
- distinct
- intersection
- cartesian
- cogroup
- sortByKey
- aggregateByKey
- repartition
- partitionBy
- coalesce
- pipe
- partitionBy
- mapWith
- countByKey
- foreach...

Keeping it simple: SQL for common operations on structured data

- Optimized execution by query processing engine

Seamlessly mix SQL and higher-order functions

- Within the same Scala, Java or Python Spark application
Machine Learning
Machine Learning

In a recent MIT study, respondents evaluated use cases for machine learning:

- 76% used machine learning to target higher sales growth
- 40% used them to improve sales and marketing performance
- 10% used machine learning to increase product sales and reduce churn.

Enterprises are using machine learning to better serve their customers with higher relevance.

Machine Learning models need to scale and that is where the power of Cloudera Enterprise excels.

**Source: Forbes Online Machine Learning Is Redefining The Enterprise In 2016**
Apache Spark MLlib
Collection of mainstream machine learning algorithms built on Spark

Including:

- **Classifiers**: logistic regression, boosted trees, random forests, etc
- **Clustering**: k-means, Latent Dirichlet Allocation (LDA)
- **Recommender Systems**: Alternating Least Squares
- **Dimensionality Reduction**: Principal Component Analysis (PCA) and Singular Value Decomposition (SVD)
- **Feature Engineering & Selection**: TF-IDF, Word2Vec, Normalizer, etc
- **Statistical Functions**: Chi-Squared Test, Pearson Correlation, etc
Real Time Analysis
Spark Streaming
Real-time and continuous processing of data streams

• Fault-tolerant and high-performance processing of continuous streams of data
  • High throughput with sub-second latency

• Similar API and programming paradigm for batch and stream processing
  • Express complex processing logic on data streams
  • Focus on the processing logic, instead of stream topology
  • Re-use code across batch and streaming jobs

• Simplified APIs for common streaming tasks:
  • Operations on “Rolling Windows”
  • Maintain and update arbitrary state for streaming events
  • Incremental aggregations

• Combine with MLlib for Predictive Analytics on streaming data
Spark Adoption

64% of current adopters plan to increase Apache Spark usage over the next 12 months.

Spark in the Cloud
Today = 23%  Future = 36%

Spark deployment in public cloud is projected to increase from 23% today to 36% in the future.
Spark in the Cloud
Why Cloudera for Spark in the Cloud?
Rely on the most portable, cost-effective, cloud-ready data platform

**Flexible Deployment**
- No vendor lock-in
- Multi-cloud and on-prem
- Transient and long-running clusters
- Flexible cluster topologies

**Flexible Pricing**
- Pay-as-you-go cloud usage
- Traditional node-based licensing
- Spot instance support
- Grow/shrink clusters

**Integrated Data Platform**
- Build end-to-end data apps
- Ingest, process, explore, model, analyze, serve
- Common security, governance, metadata, management

**Cloud-Native**
- Direct Spark I/O from S3
- Data/metadata persistence across cluster lifecycles
- Fast self-service clusters
- Single pane of glass for multi-cluster view
**Data Engineering and Data Science**

**Two Common Workload Patterns**

**Batch Processing / ETL**
(also: Testing Environments)

- Only pay for what you need, when you need it
- Transient clusters
- Single user
- Sized to demand
- Object storage centric
- Cloud-native deployment

**Exploratory Data Science**
(also: Development Environments)

- Explore and analyze all data, wherever it lives, on demand
- Transient or persistent
- Single or multi-user
- Elastic workload
- HDFS or object storage
- Lift-and-shift or cloud-native deployment
Spark in the Cloud
Sample Architecture

Kafka + Spark Streaming on permanent clusters, for streaming data ingest and processing.

Spark batch jobs on transient clusters, for processing or machine learning, directly read/write to the object store.

Interactive Spark or Impala for exploratory data science on permanent or transient clusters, directly read/write to the object store.

Serving tier (e.g. HBase, Search) on permanent clusters, serving data to end applications.

Object Store
Spark 2.0

What’s New?
New unified API: Dataset API

**RDDs**
- Object Oriented
- Functional Operators
  - map, reduceByKey, cogroup, etc
- Compile-time Type Safety

**Dataframes**
- Structured
- Compact binary representation
- Query Optimizer
- Sort/shuffle without deserialization

Datasets
Continued Innovation: Structured Streaming

Spark Streaming 2.0

- Streams modeled as continuous Dataframes
- SQL like syntax to author stream processing
  - Open stream processing to a wider audience
  - With a wide array of in-built aggregation and statistical functions
- Easier end-to-end exactly once semantics
- Out-Of-Order data handling
- Increased performance
- Growing array of Streaming ML functionality
Continued Innovation: Machine Learning Persistence

Save and Load Models

Save and Load Pipelines

- Bag of words
- Tokenize
- TF-IDF
- LDA
- Scale & Normalize Features
- Train Classifier

*Sequence is repeated during Training and Scoring

**Hyper-Parameter Tuning → Repeat Sequence with different parameter values
How do I get Spark 2.0?

Download our parcel at
http://www.cloudera.com/downloads/spark2/2-0.html

Read more at
Recommended Training for Spark Users

Apache Spark Developer Training
Cloudera University’s three-day Spark course enables participants to build complete, unified big data applications.

Data Science at Scale with Spark and Hadoop
Spark and Hadoop are transforming how data scientists work by allowing interactive and iterative data analysis at scale.

Introduction to Machine Learning
The course provides an introduction to Machine Learning, including coverage of collaborative filtering, clustering, classification, algorithms, and data volume.
Thank You