Apache Spark and z Systems

Mythili Venkatakrishnan
Technology and Architecture Lead, z Systems Analytics, IBM STSM

Enabling key analytics on z Systems

Spark on z/OS offer opportunities that can yield clients cost-removal and time-to-value improvements

- Industry Industry-wide big data theme: Bring compute to data
- Portable modern skills across all platforms (Python, Java, Scala, R, JavaScript)
- IBM embracing and extending open source technology for use on z/OS platform
- IBM Spark differentiation: Data optimization, Reference Architecture and Ecosystem of Tools including query generation aids for Spark

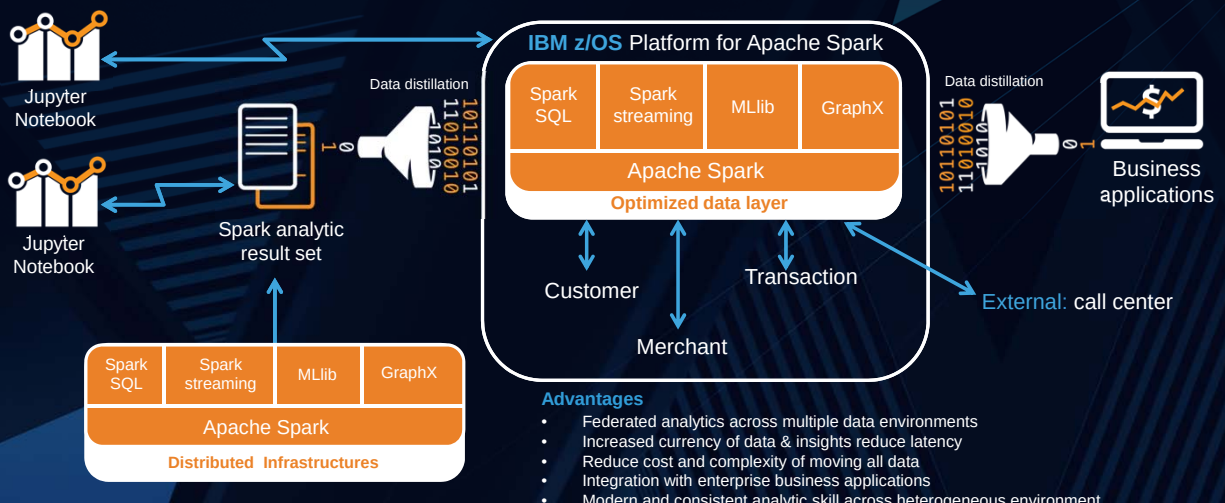
Moving all data to a data lake for analytics can yield costly side-effects

Challenges
- Development costs
- Many disparate data sources
- Data origin size and concurrency
- Ad-hoc analytical needs (schemas, freshness, etc)

Challenges
- Data Lake Management
- Security and governance issues for sensitive data
- Data and Analytic Latency
- Longer ROI for Analytics

IBM

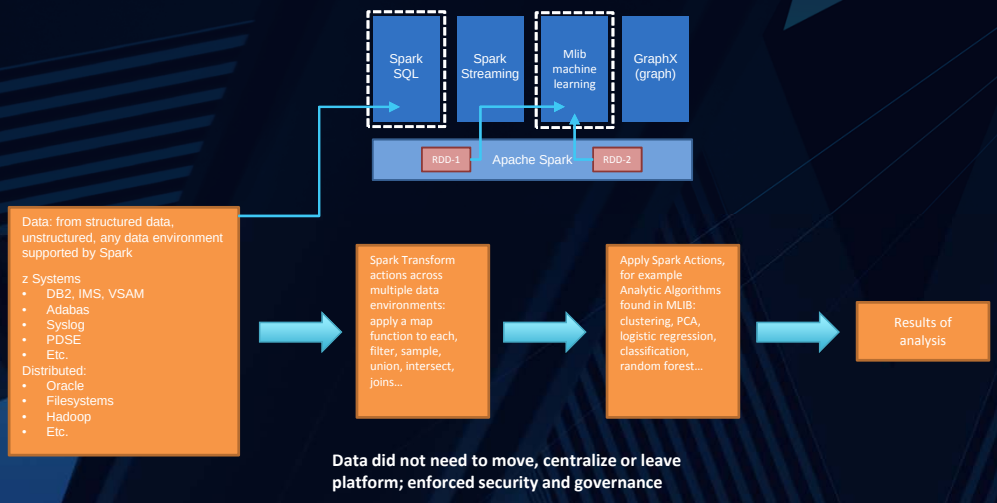
Analytic agility with Apache Spark z/OS



Advantages
- Federated analytics across multiple data environments
- Increased currency of data & insights reduce latency
- Reduce cost and complexity of moving all data
- Integration with enterprise business applications
- Modern and consistent analytic skill across heterogeneous environment

IBM

Apache Spark: Data-in-place analytics



Data did not need to move, centralize or leave platform; enforced security and governance

GA 3/25: IBM z/OS Platform for Apache Spark

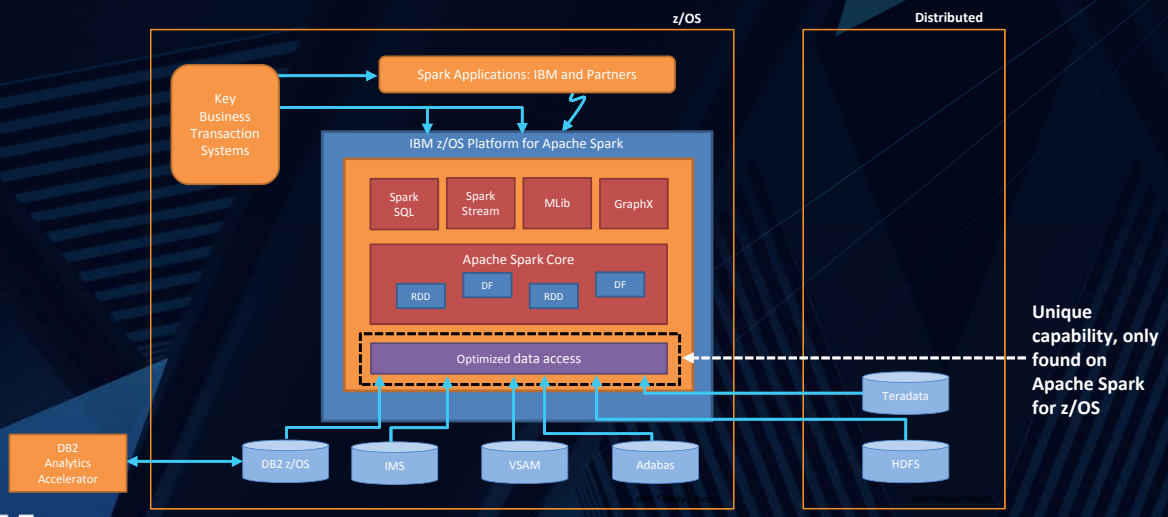
What is the offering?

- IBM z/OS Platform for Apache Spark (IBM product):
  - Apache Spark enabled for z/OS
  - Optimized Data Integration Layer
  - No License Charge product
  - Support & Service available from IBM for a fee
- Very aggressive pricing for zIIPs and memory for Spark z/OS workload
- Quick Start PoC Services available without charge – install, config, tune, data science , business solutions

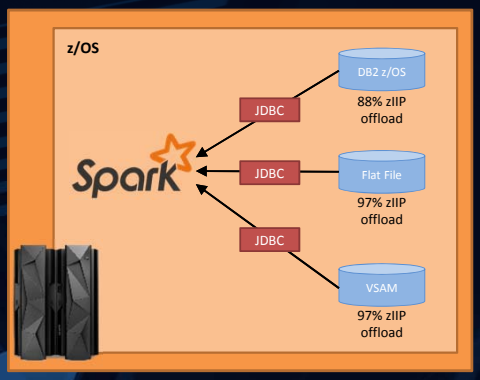
Ecosystems

- GitHub zos-spark repository
  - Jupyter Notebook IDEs (Scala Workbench, Interactive Insights Workbench)
  - Apache Job Server
  - Sample data & code snippets
- Rocket:
  - Industry vertical mappings, e.g. ISO8583-1 for card data
  - In progress: “R” support
- DataFactZ:
  - Custom Solutions for banking & insurance
- Zementis:
  - Fast, Scalable, In-Transaction Predictive Scoring integration Apache Spark

z Systems and Apache Spark



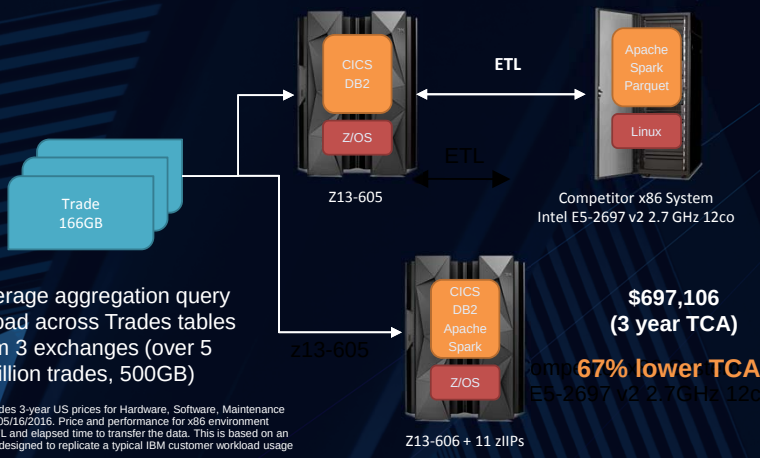
Spark on z/OS – Powerful, fast analytics without moving data



- Spark on z/OS joins multiple data types for fast, complete analytics, without moving the data
- Test of >350M rows read, parsed, analyzed, and summarized (approx. 60gig)
- Average Spark processing times – average of 3 minutes on a single Z13 LPAR with 1 GP, 13 zIIPS and 512Gb memory:
  - DB2: 2.35 minutes (4.1 mins. maximum)
  - Flat File: 2.95 minutes (3.2 mins. Maximum)
  - VSAM: 2.80 minutes (3.3 mins. Maximum)

Use Case: Large Data Pull --- bring back all 350Million rows from *each* data source, touch each data element and run Spark aggregation across all data

On-platform operational analytics achieves 67% lower TCA



$2,105,990
(3 year TCA)

$697,106
(3 year TCA)

67% lower TCA*

Brokerage aggregation query workload across Trades tables from 3 exchanges (over 5 Billion trades, 500GB)

* 3-Year TCA includes 3-year US prices for Hardware, Software, Maintenance and Support as of 05/16/2016. Price and performance for x86 environment includes cost of ETL and elapsed time to transfer the data. This is based on an IBM internal study designed to replicate a typical IBM customer workload usage in the marketplace.

It doesn't pay to move data to x86 to run analytics beyond 150GB