Benefits of Apache Spark on z Systems

George Wang

IBM

Session Code: E10
Agenda

• IBM z Systems:
  • Current capabilities and future plans on DB2 and IDAA

• What’s Spark strategy
  • What is Spark
  • Why Spark matters
  • IBM z Systems and Spark

• Incorporation of Spark for analytics on DB2 data
  • Spark project on XML analytics
  • Demo

• Q&A
Disclaimer

Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision. The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.
Businesses must evolve or be marginalized

Most Businesses Are Here

Big Data Maturity

Value

Operations

Lower the Cost of Storage

Data Warehousing

Warehouse Modernization
- Data lake
- Data offload
- ETL offload
-Queryable archive and staging

Data-informed Decision Making
- Full dataset analysis (no more sampling)
- Extract value from non-relational data
- 360° view of all enterprise data
- Exploratory analysis and discovery

Line of Business and Analytics

New Business Imperatives

Business Transformation
- Create new business models
- Risk-aware decision making
- Fight fraud and counter threats
- Optimize operations
- Attract, grow, retain customers
DB2 12 for z/OS: Redefining enterprise IT for digital business and the mobile app economy

- Scale and speed for the next era of mobile applications
  - Over 1 Million Inserts per second measured, will scale higher
  - 256 trillion rows in a single table, with agile partition technology

- In Memory database
  - 23% CPU reduction for lookups with advanced in-memory techniques

- Next Gen application support
  - 360 million transactions per hour through RESTful web API

- Deliver analytical insights faster
  - Up to 2x speedup for query workloads, 100x for targeted queries

Beta begins March, 2016
Approximately 20 customers
DB2 12 Significantly Faster Query Processing

- SAP Fiori
- SAP SFIN
- WAS Portal...
- SAP/BW
- Cust1-Uncl
- Cust1-Clus
- TPCH-SQL/PL
- TPCH-seq
- TPCH-parallel
- TPCD
- Cust2
- BIDAY-Long
- Cust3
- Crystal...
- BIDAY-Short

CPU time %

0% 5% 10% 15% 20% 25% 30% 35% 40% 45% 50% 55% 60% 65% 70% 75% 80% 85% 90%
Hybrid transaction/Analytical processing

Transaction Processing

Analytics Workload

The hybrid computing platform on z Systems

Supports transaction processing and analytics workloads concurrently, efficiently and cost-effectively

Delivers industry leading performance for mixed workloads

The unique heterogeneous scale-out platform leads in the industry

Superior availability, reliability and security

DB2 Analytics Accelerator and DB2 for z/OS

A self-managing, hybrid workload-optimized database management system that runs every query workload in the most efficient way, so that each query is executed in its optimal environment for greatest performance and cost efficiency
z13 the world’s premier data and transaction server

141 high performance cores delivering 40% more capacity

320 channels Each delivering up to 63% more data per second

Simultaneous multi-threading for up to 30% more throughput

Crypto Express5S accelerating speed of encryption up to 2x

10TB memory to eliminate I/Os for up to 70% faster response time

Up to 61% faster DB2 transactions working with the DS8870 and zHyperWrite

Now empowering the mobile generation
z/OS Data is important for Hadoop / Spark Analytics

Transaction and Log Data are consistently PART OF the analysis!

Evidence of the value of combining traditional data with Hadoop analytics

Moving Hadoop / Spark processing closer to the data on z Systems can be advantageous

What is Spark?

- An Apache Foundation open source project; not a product
- An in-memory compute engine that works with data; not a data store
- Enables highly iterative analysis on large volumes of data at scale
- Unified environment for data scientists, developers and data engineers
- Radically simplifies the process of developing intelligent apps fueled by data
Apache Spark is...

Fast
- Leverages aggressively cached in-memory distributed computing and JVM threads
- Faster than MapReduce for some workloads

Ease of use (for programmers)
- Written in Scala, an object-oriented, functional programming language
- Scala, Python and Java APIs
- Scala and Python interactive shells
- Runs on Hadoop, Mesos, standalone or cloud

General purpose
- Covers a wide range of workloads
- Provides SQL, streaming and complex analytics

from http://spark.apache.org
Brief History of Spark

2002 – MapReduce @ Google
2004 – MapReduce paper
2006 – Hadoop @ Yahoo
2008 – Hadoop Summit
2010 – Spark paper
2011 – Hadoop 1.0 GA
2014 – Apache Spark top-level
2014 – 1.2.0 release in December
2015 – 1.3.0 release in March
2015 – 1.4.0 release in June
2015 – 1.5.0 release in September
2016 – 1.6.0 release in January
2016 – 1.6.1 release in March

• Spark is Active
• Most active project in Apache Software Foundation
• One of top 3 most active Apache projects
• Databricks founded by the creators of Spark from UC Berkeley’s AMPLab
Why does Spark matter to a business?

1. Spark makes it easier to access and work with all data
   - Enables new data-based use cases
   - All data: Internal / External, Structured / Unstructured
   - Real-time insights, from all data sources
   - Automates analytics with machine learning
   - Clients that lead in data, lead in their industry

2. Spark lets you develop line-of-business applications faster

3. Spark learns from data and delivers in real-time
Who uses Spark

Build models quickly. Iterate faster. Apply intelligence everywhere.

Data Engineer
- Put right data to work for the job at hand
- Abstract data access complexity
- Enable real-time solutions

Application Developer
- Build analytics applications
- Optimize performance
- Leverages machine learning embedded

Data Scientist
- Identify patterns, trends, and risks
- Discover new actionable insights
- Build new models
  https://datascientistworkbench.com
Spark processes and analyzes data from ANY data source

Business Applications and Business Intelligence

Apache Spark

Spark SQL
Spark Streaming
MLlib (machine learning)
GraphX

Hadoop
Database
Mainframe
Data-warehouse
Resilient Distributed Datasets (RDDs)

- Spark’s basic unit of data
- Immutable, fault tolerant collection of elements that can be operated on in parallel across a cluster
- Fault tolerance
  - If data in memory is lost it will be recreated from lineage
- Caching, persistence (memory, spilling, disk) and check-pointing
- Many database or file type can be supported
- An RDD is physically distributed across the cluster, but manipulated as one logical entity:
  - Spark will “distribute” any required processing to all partitions where the RDD exists and perform necessary redistributions and aggregations as well.
  - Example: Consider a distributed RDD “Names” made of names

<table>
<thead>
<tr>
<th>Partition 1</th>
<th>Partition 2</th>
<th>Partition 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Michael</td>
<td>Cindy</td>
<td>Dirk</td>
</tr>
<tr>
<td>Jacques</td>
<td>Dan</td>
<td>Frank</td>
</tr>
<tr>
<td>Dirk</td>
<td>Susan</td>
<td>Jacques</td>
</tr>
</tbody>
</table>
Common, popular methods to access data

- **Spark SQL**
  - Provide for relational queries expressed in SQL, HiveQL and Scala
  - Seamlessly mix SQL queries with Spark programs
  - Provide a single interface for efficiently working with structured data including Apache Hive, Parquet and JSON files
  - Standard connectivity through JDBC/ODBC
  - Spark z/OS provides unique functionality to access data sources that do not natively support JDBC, as well as ability to include SQL92 and SQL99 constructs

- **Spark R**
  - Spark R is an R package that provides a light-weight front-end to use Apache Spark from R
  - Spark R exposes the Spark API through the RDD class and allows users to interactively run jobs from the R shell on a cluster.
  - Goal to make Spark R production ready
  - Rocket Software has announced intent to support R on z/OS
Spark and Hadoop

- Spark analytics *can* use information from Hadoop based file systems – though not required

- Hadoop’s MapReduce processing is disk-dependent; Spark relies on in-memory capabilities; this can result in more ‘real-time’ capabilities for Spark

- Spark comes with a set of libraries and functions for streaming, SQL, graph, and machine learning; Hadoop has an ecosystem of other tools to support these (Storm, Hive, Giraph, Mahout) which are integrated separately

- Spark can use HDFS, but is not limited to particular data store format
Spark and z Systems

- Spark processing of z data
  - Access to z data
    - JDBC access to DB2 z/OS, IMS ✔
    - Rocket Mainframe Data Service for Apache Spark ✔
- Spark running on z Systems
  - z/OS ✔
  - Linux on z Systems ✔
    - Direct download or IBM BigInsights V4.1
  - Accelerator, i.e., IDAA – future direction

Downloads

https://www.ibm.com/developerworks/java/jdk/spark/
IBM Open Platform with Apache Hadoop
Adopts Open Data Platform initiative (ODPi)

- BigInsights will include ODPi certified Apache packages
  - ODPi will initially target core packages of a Hadoop distribution
  - Packages will expand over time
  - First certification set this summer

- Our goal for BigInsights on ODPi
  - Better compatibility and less testing against ecosystem software
  - Enable IBM Hadoop capabilities to run on other ODPi-certified Hadoop distributions
IBM BigInsights for Apache Hadoop

Apache Spark is now available on both Linux on z Systems, and IBM z/OS platform

Intel Servers | IBM Power | IBM z Systems | On Cloud

Your choice of infrastructure and deployment model
IBM z Systems & Apache Spark:
Strategic Direction for Federated Analytics

- Unified Analytics Platform
- Flexibility & Agility with multi-language support
- Efficient Structure – 100x vs. in-memory map reduce
- Rich set of built-in functions with consistent APIs: Spark SQL, Spark Streaming, GraphX, ...

Leverage Linux on z Systems benefits

Leverage z/OS data and transactions

Leverage call center, external, social, sentiment data...
Apache Spark z/OS
Available since Year End 2015 via Open Source

Securely Integrate OLTP and Business Critical Data
Integrate:
- DB2 for z/OS, IMS, VSAM, PDSE, Syslog, SMF, ...
- Remote (non-z) data on distributed servers, Hadoop, Oracle, ...

NEW!
Partnership with Rocket Software supporting Apache Spark z/OS

Apache Spark Core
- Spark SQL
- Spark Stream
- MLlib
- GraphX

Spark Analytic Applications:
- Customer Provided
- IBM Provided
- Vendor Provided

OLTP Applications
CICS, IMS, WAS ...

Analysis routines

Java z/OS

DB2 z/OS

IMS

VSAM

SMF

And more...
What Makes Sense for Your Environment?

Hadoop / Spark on the Mainframe

- Data originates mostly on the mainframe (Log files, database extracts, other files)
- z governance & security models important
  - Data security a primary concern
  - Data should not be sent over the network
- Hadoop value mainly for richness of tools

Hadoop / Spark off the Mainframe

- Most data originates off of the mainframe
- Hadoop is valued for its ability to economically manage large datasets
- Desire to leverage inexpensive processing and potentially cloud elasticity
DB2 as Spark Data Source

zIIP eligible on z/OS for low cost resource consumption
**SQL on Open Source and Data in DB2 for z/OS**

*Leveraging Spark SQL Connectivity*

- **Task**: e.g. generating BI-type of reports
- **Data sources**: in DB2 for z/OS and open source data stores, i.e.
  - HDFS, Hive, Cassandra
- **Role**: e.g. Finance Manager
- Similar to „SQL on Hadoop“, leveraging Spark SQL and Big SQL
- Requires JOINs and federation

---

**Spark SQL**
- serves as entry point

**DB2 for z/OS**
- Spark SQL
- JDBC Access
- RDDs
- DataFrames
- Other data sources possible

**Processing RDDs**
- RDD Method
- Creating DataFrames

**Spark SQL serves as entry point**
- Works today
- Optimization work currently done

**Creating DataFrames**
- (optional)
- DB2 Analytics Accelerator

---

**Other data sources possible**
- HDFS, Hive, Cassandra
- Cassandra
IMS as a Spark Data Source

Available Now!

IMS JDBC Type-4 Connection
IMS JDBC
IMS Connect
IMS ODBM

IMS JDBC Type-2 Connection
IMS JDBC
Native
Summary

• Spark is of increasing strategic importance to our enterprise customers
  • Don’t think that Spark is seen relevant and implemented purely in “distributed”

• Moving Spark processing to close proximity of z data can provide many benefits
  • Several Options are already available and more are planned

• Spark brings many new opportunities for opening up new analytics capabilities for your valuable z Systems data assets
  • SQL users as well as Non SQL users
Incorporating Spark Analytics using z Data Sources for Large Scale XML Data Processing (Project Background + Demo)
Introduction

• Objective
  • Enabling business data analytic support without impacting OLTP
  • Enabling real-time analytics on XML data for hotspot, fraud detection, etc.

• Approach
  • Middleware-based data transfer for capturing XML data into Spark memory only for efficiency
  • Efficient data layout design for enabling query support on XML data
  • Spark for advanced analytics capability using Spark SQL
XML Challenges

- Parsing XML is CPU intensive
- XML in database not feasible to aggregate/analyze the data
- Complicated query on large document parsing result in more memory consumption
- Large and deeply nested XMLs makes problem worse
- Special handling of optional fields
- Repeated parsing for each query when scanning the table causes high I/Os
- Potentially a storage issue
IBM DB2 Analytics Accelerator

CLIENT

Data Studio with DB2 Analytics Accelerator Studio Plug-in

z Systems

Dedicated highly available network connection

PureData System for Analytics (Netezza Technology)

Users/Applications

Data Warehouse application DB2 for z/OS enabled for IBM DB2 Analytics Accelerator

IBM DB2 Analytics Accelerator

IBM Systems

IBM Systems

IBM Systems
Solution with Spark

- Use of IBM InfoSphere BigInsights for Apache Spark
  - Heavy lifting data processing using Spark framework
  - Spark engine to scan the table for analytic and aggregation
- Middleware support
  - Initial data streaming from DB2 via JDBC/RDD connection driver to Spark memory, HDFS is not required!
  - Parallel data fetching from user source tables to Spark memory
- Analytic pipeline
  - XML Parsing
  - Mapping for data record normalization
  - Transform records in DataFrame
  - Spark SQL
- Cloud-based data exploration using Zeppelin
  - data visualization
Spark on z Data

IBM Open Platform (IOP) for Advanced Analysis

- **DataFrames**
  - Spark SQL
  - Spark Streaming

- **ML Pipelines**
  - MLlib
  - GraphX

**Spark Core**

- **Analytic data pipelining**
- **Result reporting**

- **OLTP workloads**
- **OLAP query workloads**

**Mobile Analytic Solution**
Data Processing by Spark

- **Input Data**
  - Flat Array in Spark memory

- **RDD(1)**
  - Memory Resident
  - Parsing Data

- **RDD(1)**
  - Not Memory Resident
  - Parsing data

- **RDD(2)**
  - Memory Resident
  - Data mapping iteration

- **RDD(N)**
  - Memory Resident
  - ETL transformation

- **Transform an RDD to a DataFrame**

- **Spark executor on SQL statement**

- **Can be swapped into disk or recreated on read**
Demo: OECD Statistics Analysis

Demo Story: OECD (Organisation for Economic Co-Operation and Development)

Matrices showing employment rates of European countries (quarterly), by gender is loaded into XML table and initial data pipelining into Spark memory

http://db2bigxml03.svl.ibm.com:8069
1) XML is transferred from DB2 as text representation through the network to the JDBC RDD

2) The text is parsed on the Spark side into an XML document tree

3) XPath is used to find and extract the values from XML

4) The extracted values are stored in simple case objects, where each object represents one entity

5) A DataFrame is defined across the stream of case objects, making the data accessible to SparkSQL

6) Data Scientist can use SparkSQL, Notebooks or visualization tools to query and analyze the data
Spark z/OS Demo: Configuration

Cloudant
Linux on z

REST API

CICS Transaction system

DB2 z/OS

IMS

IMS Transaction system

VSAM

JDBC

Spark Job Server

Open Source Visualization Libraries

Credit Card Info

Trade Transaction Info

Customer Info

© 2014 IBM Corporation
SHOW DEMO LIVE OR LINK:

HTTPS://YOUTU.BE/SDMWCUO5RK8
IBM Spark Technology Center (STC)

- Focal point for IBM investment in Spark
  - Code contributions to Spark project
  - Build industry solutions using Spark
  - Evangelize Spark technology inside/outside IBM

- Agile engagement across IBM divisions
  - **Systems**: contribute enhancements to Spark core, and optimized infrastructure (hardware/software) for Spark
  - **Analytics**: IBM Analytics software will exploit Spark processing
  - **Research**: build innovations above (solutions that use Spark), inside (improvements to Spark core), and below (improve systems that execute Spark) the Spark stack
George Wang
IBM
fqwang@us.ibm.com

E10
Benefits of Apache Spark on z Systems

Please fill out your session evaluation before leaving!