Derivatives Analytics using Distributed Computing on Big Data Platform

By

Khader Shaik

Agenda

- Derivatives Analytics
- Current Systems' Challenges
- Big Data Quick Review
- Distributed Computing & Big Data Platform
- Architecture for Derivatives Analytics
- DAG for Computing Analytics
- Platform Tools
- Q&A

Derivatives Analytics

- Key Computations – Pricing, P&L, Risk Analytics and Simulations
- Critical for Trading, Hedging decisions and Risk Management
- Key Areas
  - Quantitative research
  - P&L Analysis
  - Risk Mgmt. (Market Risk, Credit Risk)
  - Limits Management
  - Collateral Optimization
  - Regulatory Reporting
- Computation Frequency
  - Batch - End of Day, Weekly; Monthly; Quarterly; Annual
  - Near Real-time - intra-day on-demand

Current Systems' Challenges

- Most Sell-side firms develop vertical systems by asset classes or desks
- Buy-side firms typically rely on vendor products for analytics
- Common issues of current systems
  - Mostly batch oriented
  - Require large amount of resources – quite costly and long running processes
  - Not easily scalable
  - Less flexible (rigid architecture) – hard to add new/custom models
  - Current out-of-the-box solutions are not so flexible - hard to add new/custom models
  - Many firms are do not have expertise build analytics systems
  - Limited vendors support multi-asset class products
  - Most of them are not yet ready to use Distributed computing techniques (some are already migrated, others are joining)

Big Data

- Big Data – Volume, Variety, Velocity
  - Unstructured and structured data generated from diverse sources in too large volumes
  - Not possible to handle using traditional technologies
  - Social media, news feeds, blogs, internal and other sources
- Big Data Analytics
  - Analytics applied to Big data in order to add value to business
  - Examples – detect business opportunities, improve operational efficiencies, develop targeted marketing campaigns and detect risks

Big Data Evolution

- Big Data Evolution includes two parts
  - Big Data Analysis – data analysis models and measures that can be used to add value to business
  - Big Data Platform – tools and technologies used for Big Data analytics, which are lot faster and cheaper than traditional technologies

Big Data - Use Cases

- Ideal in situation where
  - Extremely large amount, unstructured and speed of data involved
  - Can also be used in other situations like complex and resource intensive processing that can't be handled by traditional technologies
- Large-scale users of Big Data platform
  - Google, Yahoo, Facebook, Twitter, LinkedIn, Walmart and many more
  - Also Pharmaceuticals for research, Banking for fraud detection, risk mgmt., and marketing

Foundation of Big Data Platform

- Built-in distributed computing and distributed storage
- Utilizes multiple resources in parallel – storage, and processing
- Runs on cluster of machines made up with cheap commodity hardware (Grid) or Cloud (multi-core)
- Built in fault-tolerance

Designing Systems on BD Platform

- Big Data platform is quite matured
- Some large- and mid-size financial institutions have already started using
- Some institutions are still trying to find how and where to use it right
- Identify right business problems that require BD solution - *Right solution for right problem*
- Right architecture is critical to leverage power of BD platform
- First focus on architecture not tools (some firms simply started developing apps using BD technology)
- Do not simply replace current RDBMS or other application technologies with BD technology

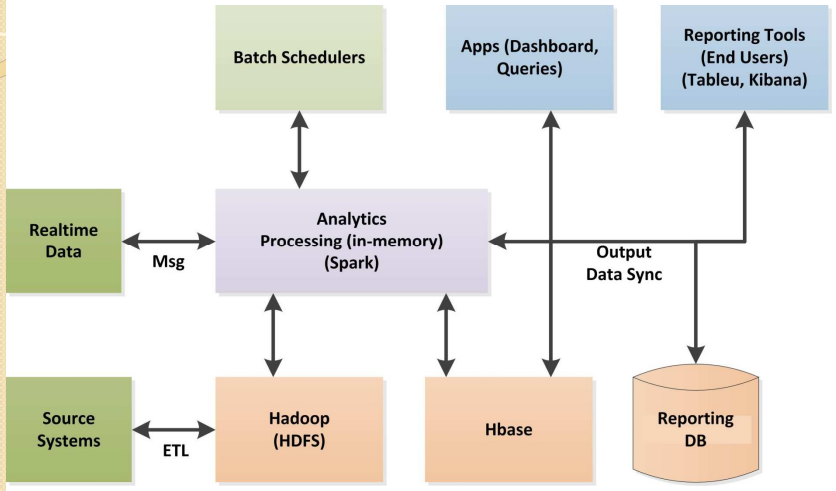
Distributed Computing on Grid

- Next generation and advanced computing environment
- Provides real-time or near real-time performance
- Major differentiator from traditional platforms
- This architecture combines the core Big Data tools and advanced powerful and faster computing tools
- Ideal for systems that involve large data and complex processing in batch environment
- BD can even serve real-time (near) needs with latest tools and better architecture
- Perfect candidate for derivatives analytics, enterprise risk management, regulatory reporting, collateral optimization, pre-trade analytics, research solutions

Distributed Computing Architecture

- Highly Scalable (scale out instead scale up) - Easy to add processing power on demand
- Fault-tolerant - In-built fault tolerance features
- Low cost - Cheaper commodity hardware and most tools are open-source
- Faster time-to-market solutions
  - Many stable and powerful open-source software tools & libraries
  - Involvement of large and active open-source community involved
  - Dedicated support from multiple commercial firms (Eg. Cloudera, Hartonworks)

Big Data Architecture for Derivatives Analytics (BigDADA)



Advanced Architecture - Features

- Essentially based on Lambda Architecture
- Combines batch and (near) real-time processing
- Uses DAG for model building
- Uses in-memory computation layer which drastically increases speed
- Built on Big Data Platform
- Uses Open-source tools and technologies

Features ..cont

- Better utilization of Cloud, Cluster or Grid
- Near-real time results (faster processing)
- Compute only impacted metrics and data points - Directed Acyclic Graph (DAG) model
- Fast scenario analysis, pre-trade analysis and decision making
- Seamless integration of new models
- Parallel model computations - easy model validation
- Support for large portfolios and multi-asset class products
- Live data visualization capabilities (Integrated Tools such as Tableu, Kibana)

Layers & Key Components

- Data Layer
  - Hadoop Platform - HDFS (unstructured) , HBase (Semistructured), Reporting Database for (BI) Analysis tools
- Source Data Import Layer
  - ETL and other tools – simple tools to load data from various source systems
- Real-time Data Load
  - Data from real-time sources using messaging (Kafka or other)
- Computation Layer
  - In-memory computation using cache – Spark or similar tools
  - DAG model for coding financial models
  - Financial Libraries from vendors with support for distributed computing (grid ready)
  - Stream ??
- Presentation Layer - Tableu, Kibana, Excel, APIs (for custom apps)

DAG

- Directed Acyclic Graph (DAG) or Acyclic Digraph – is a directed graph with no directed cycles. That means, there is no path that starts at one node and comes back to same
- It helps to model computations that run in parallel and require re-computes of only impacted components of a formula
- Ideal to combine batch and real-time analytics. Delivers results faster after first compute.
- Speeds up intra-day risk analysis, simulations, P&L Analysis, Trade decision support

Financial/Quantitative Libraries

- Options – build grid-ready models using standard libraries or license vendor models
- All models must support underlying distributed computing (grid)
- Python – is a popular and most suitable language to develop in-house libraries
- Mathematical libs are - NumPi, SciPi, Pandas and other
- It is also possible to migrate existing custom models if any
- Some commercial libraries are already grid ready

Popular Opens Source Tools

- Data Layer – Hadoop, MapReduce,, Hbase Cassandra
- Processing – Hive, Spark, MapR – in-memory distributed computing (near-data processing)
- Messaging - Kafka (distributed messaging)
- Flume – Data transfer of big data
- Elasticsearch, Kibana
- Presentation Layer - Tableu, Kibana, Excel, APIs (for custom apps)
- Financial Libs - Python, R or any financial libraries

Takeaway

- Hadoop storage is not replacement for your current RDBMS. Also, it is not replacement for EDW/ETL
- Big Data platform is ideal for specific purpose such as to compute analytics in batch or near real-time. Originally, Hadoop platform is not built for real-time.
- Advanced architecture could achieve high performance distributed computing (parallel/grid computing) – achieve speed, efficiency and cheap
- Big Data platform could be a complement to existing infrastructure and help achieve in fraction of time and cost. Also enables processing of unstructured data.

Q&A

Thank You

Khader Shaik

www.ksvali.com