Challenges of Managing Risk and Regulatory Data and Big Data Solution

By
Khader Shaik

Risk USA - 2016
“It’s easy to lie with statistics. It’s hard to tell the truth without statistics.”
– Andrejs Dunkels
Agenda

- Regulatory Environment & Data Challenges
- Data Governance Principles
- Unified Data Platform Solution
- Overview of Big Data
- Big Data Platform Features & Financial Institutions
- Big Data Unified Platform for Data Governance and Risk and Regulatory Data
- Big Data Adoption Roadmap & Strategy
- Takeaway
- Questions & Answers
Regulatory Environment & Data Challenges
Regulatory Initiatives

- Dodd-Frank
- EMIR
- Basel III
- BCBS 239
- CCAR
- Solvency II
- MiFID II
- FATCA
- COREP/FINREP
- ACORD
- ..Other
Key Ideas behind Regulations

- Financial Stability and Transparency
  - Visibility of risk exposures all the way up to enterprise level
- Central comprehensive data governance and risk mgmt. infrastructure
  - Efficient Data management
- Efficient Risk Management Processes
  - Better risk data aggregation and reporting
- Data Harmonization
  - Standardized reporting across the board
- Global regulatory standards for capital adequacy, stress testing and market liquidity risk
- Formation of various agencies for market surveillance
Challenges facing Financial Institutions

- **Transparency**
  - Need to collect and process massive amounts of data including transactions, events, reference data, market data and more into central data hub

- **Powerful Computing Infrastructure** – requires next generation technology infrastructure to:
  - Collect and store large sets of data
  - Computation of complex algorithms on large sets of data (pricing, risk models, stress tests, simulations)
  - Process in real-time or at least in near real-time and batch mode

- **Enhanced Data Governance**
  - New practices of data ownership and accountability to achieve risk data quality
  - Data lineage, accuracy, completeness, timeliness and adaptability
  - Need of data life cycle management and process flows

- **Better Risk Analysis, Aggregation and Reporting**
  - Produce reconcilable risk and reporting data

- **Bottom line** - Strategic Risk and Reporting Solution with as much importance as any trading or financial systems
Challenges of Current Architecture

Current IT Infrastructure

- Many independent systems with different objectives
- Many of them can't scale efficiently to meet today’s regulatory requirements
- Multiple Vendor or in-house, fragmented and independent risk analysis and reporting systems
- Most Sell-side firms develop vertical systems by asset classes, desks or business units
- Buy-side firms typically rely on vendor products for analytics
- Out-of-the-box vendor solutions are less flexible
- Current systems operate mostly in batch mode producing file based data
File based output doesn't allow drill-down capabilities
Decentralized, mostly file based regulatory reporting data - hard to build mgmt. dashboards
No integrated risk data that can be analyzed deeper or used to run any type of optimization
Requires large amount of resources – costly and long running processes
Not easily reconcilable data and no proper audit trail
Less flexible architecture – hard to add new or custom models
Data Management Challenges

- Data is all over the place in many different databases with no consistency in reporting of values
- Data is not clean enough to perform analysis and we can’t reconcile manually
- Each department has a different analytical objectives
- Most risk processes run in Batch mode (no real-time)
- Risk exposure aggregation is difficult to trace. Limited or no tools to drill-down or slice and dice any exposures
- Regulatory reporting data is produced in files, and the data elements are not traceable
- Hard to map and understanding data relationships
- Hard to aggregate and analyze notional amounts other metrics at enterprise level
Data Governance Principles & Risk Analytics
Data Management - BCBS 239 (RDARR)

- Basel Committee introduced Risk Data Aggregation and Risk Reporting (RDARR) a.k.a BCBS 239
- Compliance requirement for SIBs
- Introduces set of principles for risk data governance and risk data aggregation and reporting
- Very comprehensive reform and proved to be very expensive
- Many banks considered this as an opportunity to fix their risk processes
- Key objectives
  - Enhance risk data infrastructure
  - Improve the speed of producing the risk data
  - Enhance the management of information
  - Reduce the probability and severity of losses resulting from inefficient risk management
  - Improve the quality of strategic planning and its ability to manage the risk of new products and services
Benefits of Better Risk Data Management

- Improved decision-making capability by enhancing data integrity
- Create new insights to support enterprise-wide decision-making
- Consistent and easy updates of enterprise data
- Efficient use of risk resources - focus on analytics rather than data processing
BCBS 239 — Principles

Governance & Infrastructure
- Data Governance Principles
- Data Architecture & IT Infrastructure

Risk Data Aggregation
- Accuracy and Integrity
- Completeness
- Timeliness
- Adaptability

Risk Reporting
- Accuracy
- Comprehensiveness
- Clarity & Usefulness
- Frequency
- Distribution

Supervisory Review
- Review Process & Tools

© Copyright 2016 Khader Shaik, Risk USA — 2016
## Governance Processes

- Ownership & stewardship
- Data classification
- Data lineage (life cycle)
- Data profiling
- Data validation
- Data quality reporting and exceptions mgmt.

- Data cleansing
- Data standardization
- Reference data management
- Data workflow processes and tools
Data Lineage — Life Cycle

- Data Lineage is data life cycle - describes each step from origin to all processes it goes through
- Data Lineage is core of data governance
- The knowledge of data is critical for governance process, data quality, master data management and overall metadata management
- Data Lineage answers questions like -
  - Where did the data come from?
  - How did we arrive at this number?
Data Quality Principles

- **Classification** - consistent definition of risk and regulatory measures across the enterprise
- **Accuracy & Completeness** - Reconcilable with source systems
- **Timely** – quality checked reports on time
- **Comprehensive** – complete risk exposure from all risk areas of enterprise
- **Traceable** - across the data production process, embedded data lineage
- **Usable** – usable without any further transformation
- **Flexible & Adaptable** - useful for on-demand and ad-hoc analysis. Must accommodate any changes in reporting
Unified Platform Architecture
Solution — Unified Platform Architecture

Data Sources (Business Apps, External Sources)

Unified Data and Analytics Platform

Analytics Platform

Data Integration, Governance and Access Tools

Enterprise Data Platform

End User Apps (Risk View Tools, Dashboards, Enterprise Apps Reports)
Unified Platform - Key Features

- Large volumes of data storage and efficient access
  - Data Lake

- Comprehensive data management
  - Data Governance

- Powerful analytics for risk analytics computation
  - Analytics Engine/Platform

- Data analysis tools for both power users and common users
  - Reporting and data access tools
Overview of Big Data
Overview of Big Data

- Big Data – three V’s
  - Volume - Big data refers to huge data sets that are orders of magnitude larger
  - Variety - more diverse, including structured, semi-structured, and unstructured data
  - Velocity - arriving faster than a typical organization’s business data

- Big Data Insights
  - Insights are discovered patterns, derived meaning, indicators for decisions, and ultimately add value to the business
  - Real value of big data is in the insights it produces when analyzed
  - Examples – detect business opportunities, improve operational efficiencies, develop targeted marketing campaigns and detect risks in advance, decision making in trading

- Big Data Analytics
  - Analytics is a set of advanced technologies designed to work with Big data
  - Gaining intelligence by exploring data using various sophisticated quantitative methods such as machine learning, neural networks, computational mathematics, and artificial intelligence

- Big Data Technology - Hardware and Software
Big Data Components

Data  Insights  Analytics

Big Data Software Components

Cluster (Cloud)
Big Data is about *parallelization* - delivers *high performance computing 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 and high availability
- Enables applying simple and complex algorithms on large datasets
- Processes in batch, interactive, real-time and near real-time
Features of Big Data Platform ..cont.

- Higher ROI - Cheaper hardware leverages grid, cloud; many open source software components
- Provides powerful Data Platform (Data Lake)
- Provides powerful analytics platform (processing engines)
- Complex event processing capability to analyze continues data streams (provided real-time analytics)
- Availability of large set of software package and programming tools
- Serves needs that couldn’t be handled by traditional technologies
Terminology

- Grid – Cluster – Cloud
- Node
- Core
- Distributed computing - Parallel processing
- Multi-threading
- Resources – nodes (memory, storage and processor)
- Cluster manager – software that manages cluster
Distributed Processing — MapReduce Example

Input

- Big data platform
- Big data analytics
- Analytics platform data

Splitting

- Big data platform
- Big data analytics
- Analytics platform data

Mapping

- Big, 1 Data, 1 Platform, 1 Analytics, 1

Shuffling

- Big, (1,1)
- Data, (1,1,1)
- Platform, (1,1)
- Analytics, (1,1)

Reducing

- Big, 2
- Data, 3
- Platform, 2
- Analytics, 2

Result

- Big, 2 Data, 3 Platform, 1 Analytics, 2
Ideal situations
- Very large amount and multi-structured data is involved
- Where simple or complex algorithms to be applied on large sets of data
- Fast and easily scalable computing environment

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, research organizations and medical fields
Big Data for Financial Institutions

- **Data Lake/EDM - unlock the power data**
  - Capability to handle very large amounts of data - enterprise-wide transactions, events, master data, market data, risk metrics, regulatory reporting data and more

- **Real-time Analytics Platform – better decision making**
  - Real-time analysis of risk, capital optimization which add value to business
  - Study impact of market on portfolios and capital in real-time – optimize capital leverage

- **Complex Simulations**
  - Ability to compute complex stress tests, simulations and analyze results in near-real-time
  - Current proprietary grids being used for Monte-Carlo can also be replaced (many large firms)

- **Extensive Stress Tests**
  - Ability to execute extensive stress tests that add value to enterprise risk mgmt. process

- **Yields robust Risk & Regulatory Technology Platform**
  - Enables the firm go beyond regulatory reporting, and add infrastructure that can help improve business decision making

- **Empowers firms to add other analytics platforms**
  - Robust Sales & Marketing, End-to-end investment mgmt. systems, Trading, and portfolio mgmt. (CRM, OMS/EMS), Operations risk mgmt. – fraud detection, detecting malicious attacks etc.
Data Layer
- Hadoop Platform - HDFS (unstructured), HBase/Cassandra (Semi-structured), Reporting Database for (BI) Analysis tools

Infrastructure Management
- Cluster manager, Resource manager, Job scheduler
- YARN, Mesos, ZooKeeper and other

Source Data Integration Layer
- ETL and other tools – tools to load data from various source systems (ingest, transform and load) and apply governance rules
- Falcon, Flume other ETL and rule engines

Computation Layer
- In-memory computation engine – Spark or similar tools
- Analytical Models – self developed and/or vendor
Layers & Key Components .cont.

- Real-time Data Load
  - Data from real-time sources using messaging (Kafka, Spark Streaming, Storm and other)

- Presentation Layer Applications
  - In-house or commercial applications for data analysis and reporting
  - Tableau, Kibana, Excel, APIs (for custom apps), Platfora, MicroStrategy, QlikView and more
Platform Features

- Robust, scalable, fully integrated data management, analytics and reporting platform
- Powerful Data Management
  - Advanced EDM holding enterprise level data
  - Data Governance (BCBS 239) requirements can be complied (if applicable)
- Advanced Analytics
  - Batch, real-time or near-real-time risk data computation, regulatory data generation, other insights as required
  - All other regulatory data can be produced from a single data platform
- Parallel & Independent Analytics Platform
  - Parallel and independent data and computing platform to existing business applications
- Highly scalable
  - Big Data infrastructure scales on-demand and supports the growth easily
Comprehensive solution for enterprise data management

Full-scaled Data Governance solution
- From P&L reporting to Executive reports
- Implementation of data governance principles of BCBS 239 as required

Market Data repository – securities prices, interest rates, FX prices etc.

Reference Data repository - securities, counterparties, contract terms, etc.

Current and historical risk metrics (store & manage computed data)

Current and historical regulator reporting data (store & manage computed data)

Central & single source of enterprise data for risk and reporting
- Central repository of all positions (all types of portfolios)
- Produce all risk and regulatory reports from a single source
- Audit trail, reconciliation, data validation capabilities
Data Governance on Big Data Platform

Data Governance (Load and Manage)
- Integration Workflow System Tools (Falcon, Flume, Sqoop, etc.)

Data Platform
- Hive (SQL)
- Pig (Data flow)
- Machine Lang.
- Streaming (Kafka, Storm)
- In Memory & Other (Spark)
- MapReduce
- YARN
- HDFS (distributed file system)
- HBase/Cassandra (Columnar Database)

Data Access & Management

Resource Manager (Cluster Management, Data Processing)

Security
- Authentication Authorization
- Data Protection Accounting (Falcon, Knox etc.)

Operations
- Provision Schedule
- Manage Monitor (Ambari, Zookeeper, Oozie, etc.)
Key Computations – Pricing, P&L, Risk Analytics, Simulations, Stress tests, Regulatory data

Central engines for multiple functional analytics
- Quantitative research, P&L Analysis, Market Risk, Credit Risk, Liquidity Risk, Limits Management, Collateral Optimization, Regulatory Reporting

Real-time sensitivities computations

Real-time simulations
- Aggregation of all exposures for all trades across all desks
- Integration of new trades to help front office decision making

Slice & dice analysis across all available dimensions

Enterprise level risk exposure aggregation and limits monitoring with dashboards and drill-down capabilities

Enables implementation of better models for market risk, credit risk, liquidity risk

Ability to include complex products such as mortgage products
Data Lake maintains the copy of source data that is produced by various business applications.

Connectors load and continuously replicate changes from business application databases to Data Lake (real-time or near real-time).

Allows organization to efficiently and cost-effectively perform big data analytics.

Data can be staged in pre-processing environment. Later it can be cleansed, aggregated and loaded into high-performance analytical databases.
Analytics Platform

Applications
(Analytics, Interactive, Enterprise etc.)

Scala, Java, Python, R, APIs.  Spark SQL  Streaming  Mllib (Machine Learning)  GraphX  Other

Spark Analytics Engine

Resource Manager (YARN, Mesos or other)

Data Storage (HDFS or other)
Spark is an open-source data analytics cluster computing framework

More advanced in-memory, real-time platform

Spark can either work as part of Hadoop eco system or independent platform with other components

Speeds up intra-day risk analysis, simulations, P&L Analysis, Trade decision support

Spark is popular at top wall street firms
Spark offers a rich API for data analytics
Up to 100x faster in some applications (than MapReduce)
Supports Python, R, Java, Scala, other financial libraries
Spark coding is quit simplified
Data Access – provides access to NoSQL, SQL, and HDFS
Spark Streaming – allows processing of live data streams
MLlib – provides machine learning library
Vendor Solutions for Data Governance & Risk Analytics

- Good news, there are many commercial vendor implementations using Hadoop are available for Data Governance
  - These are open solutions, could cost effective and easier to adopt
- Also Analytics platform using Spark and required products bundled
- Grid ready mathematical and financial libraries
- Reporting and Visualization Tools
  - Quite powerful big data ready vendor products are available
- Big Data distribution vendors
  - Entire big data stack of software packages, upgrades and support
Big Data open source products are packaged and supported by various vendors.

Most vendors provide various services in addition and their own range of products.

Popular vendors are –
- Hartonworks
- Cloudera
- Databrix
- MAPR
- Cloudera
- Datastax
- Bluedata
- Oracle … more

You can also build your own platform.
Languages and Financial Libraries

- Python, Scala and Java are most favored languages in Big Data world
- Python – is a most favored language to develop in-house libraries
- Mathematical & financial libraries - NumPi, SciPi, Pandas, R, MatLab and other
Hadoop (Hadoop Common) - open source software framework that supports data-intensive distributed applications. It enables applications to work with thousands of computational independent computers and petabytes of data. It is set of common libraries and utilities used by other Hadoop modules.

Hadoop Distributed File System (HDFS) - is a distributed, scalable, and portable file system written for the Hadoop framework. HDFS allows Hadoop cluster to scale to hundreds and thousands of nodes.

MapReduce - is a software framework that executes a wide range of analytic functions by analyzing datasets in parallel before ‘reducing’ the results. The “Map” job distributes a query to different nodes, and the “Reduce” gathers the results and resolves them into a single value. It can process massive amounts of data in parallel across a distributed cluster.

Pig – is a programming platform to write MapReduce programs to query large scale data. General Pig scripts are written in the Pig Latin language.

YARN – is the cluster management layer of Hadoop. YARN replaces cluster management services of MapReduce.
Hive – is provides SQL-like interface to access to Hadoop platform

HBase – is open source, distributed, software package that provides real-time read/write access to Hadoop/HDFS

Storm - is a distributed real-time computation system. It is simple, helps process unbounded streams of data

Spark – is a in-memory distributed data processing engine. It promises speeds up to 100 times faster than MapReduce

Tez - is an extensible framework for building high performance batch and interactive data processing applications, coordinated by YARN

Oozie - is a workflow scheduler system to manage Hadoop jobs
- **ZooKeeper** - is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services

- **Ambari** – is a Hadoop cluster management software, used for provisioning, managing, and monitoring. Ambari provides an intuitive, easy-to-use Hadoop management web user interface

- **Sqoop** – is the tool designed for efficiently transferring bulk data between the Apache distribution of Hadoop and structured relational databases. (SQL to Hadoop = Sqoop)

- **Cassandra, MongoDB** – are NoSQL open source databases packages designed for scale, flexible data aggregation and to store files of any size. They are fast and support high availability and full indexing. (NoSQL query language that is not SQL, works with columnar databases)

- **Flume** - is a distributed, reliable, and high-available service for efficiently collecting, aggregating, and moving large amounts of streaming data into HDFS

.. Many more
Big Data Platform can be built on internal cluster or Cloud
- Cloud – hosted hardware platform (managed by service provider)
- Analytics services are also known as Analytics as a Service (AaaS)
- Cloud computing models can help accelerate the potential for scalable analytics solutions
- Clouds offer flexibility and efficiencies for accessing data, delivering insights, and driving value
- Choose the solution that fits your strategy – Cloud ideal for small to mid-size firms
- Options – private cloud, public cloud or hybrid
  - Private cloud to mitigate risk and maintain total control
  - Public cloud provides scalability, less resource intensive and cost-effective
  - Hybrid model combines private and public cloud resources and services
- Deciding factors are workload, cost, security, and data interoperability
Cloud Usage Models

- Infrastructure as a Service (IaaS)
  - IaaS provides just cloud hardware
  - Your organization will be responsible for installing all software required, such as the Hadoop ecosystem and managing resources
  - IaaS requires greater investment of IT resources to manage and implement big data analytics

- Platform as a Service (PaaS)
  - PaaS provides cloud infrastructure and tools and libraries to build, test, deploy, and run applications
  - PaaS reduces management workload by eliminating the need to configure and maintain infrastructure. But you are responsible to implement analytics

- Software as a Service (SaaS)
  - SaaS provides end-user applications for cloud-based big data analytics such as sentiment analysis, risk management.
  - This model is also known as pay-as-you-go

© Copyright 2016 Khader Shaik, Risk USA – 2016
Big Data Adoption Roadmap & Strategy
Big Data Adoption - Guidelines

- Make part of enterprise strategy
  - Big Data appears to be a technology initiative, but it is more than that. Make part of Enterprise Strategy

- Understand Big Data and Technologies
  - It is evolving, different and crowded, so spend time to educate

- Be prepared to change your mindset
  - Big Data projects are not typical IT projects. Big Data is to deal with big challenges

- Understand and adopt open source software model, distribution vendors, new infrastructure (cloud/grid) management
Adoption Phases

- **Data discovery** – identify and define data sources and data items for new platform
- **Analytics discovery** – identify and define analytics such as risk metrics, regulatory data
- **Tools and technology discovery** – compare and choose the software components, programming tools and APIs required
- **Infrastructure discovery** – choose and acquire Big Data platform infrastructure
- **Implementation** – Implement the solution
- **Production support** - Support and grow platform
Big Data Roadmap

- Educate on Big Data: Spend time to educate all involved
- Management Buy-in: Get the management approvals in early stage
- Update Enterprise Strategy: Update business, data, technology and delivery strategies
- Team & Infrastructure Setup: Build core team and acquire required infrastructure for development and production
- Define Use Case & Proof of Concept: Choose the problem and define the use case for proof of concept & validate
- Plan & Execute Project: Plan your project and execute it in phases
Big data demands a new set of skills, especially in IT

Developers with Big Data programming skills
  ◦ Computer Science and strong Mathematical background

Data Scientists (*New title*)
  ◦ Individuals who apply big data analytics to complex business problems and make sense of the results. *They may sit in business or IT*

Platform Administrators

Power Users – must be trained to use new set of analysis tools and language such as Python
Big Data platform not only helps you with regulatory compliance, but can also add value to your business through better (Big Data) analytics

Big Data Lake/Hadoop storage may not be a replacement for your current EDW or relational database

Today’s Big Data platform is capable of delivering real-time analytics

Total ROI could be much higher - right architecture may provide high performance, efficiency at low-cost

Big Data platform could be a complement to existing infrastructure and help build analytics in fraction of time and cost.
Big Data platform is critical for the analytics of complex products such as derivatives

Adopt data standards and governance processes for successful data management and compliance

There are too many products in Big Data world, do due diligence

Open source Big Data tools are widely used, so do not hesitate

Machine Learning can help in the areas of fraud detection, surveillance, risk model improvement, real-time analysis and alert
Questions & Answers
Thank You
Khader Shaik