PREDICTIVE ANALYSIS OF SHARES

ABSTRACT

Big data is a catchphrase for a new way of conducting analysis. Big data principles are being adopted across many industries and in many varieties. Has your Investment Management firm started implementing Big-Data?

Swarup Mishal
Author

Table of Contents

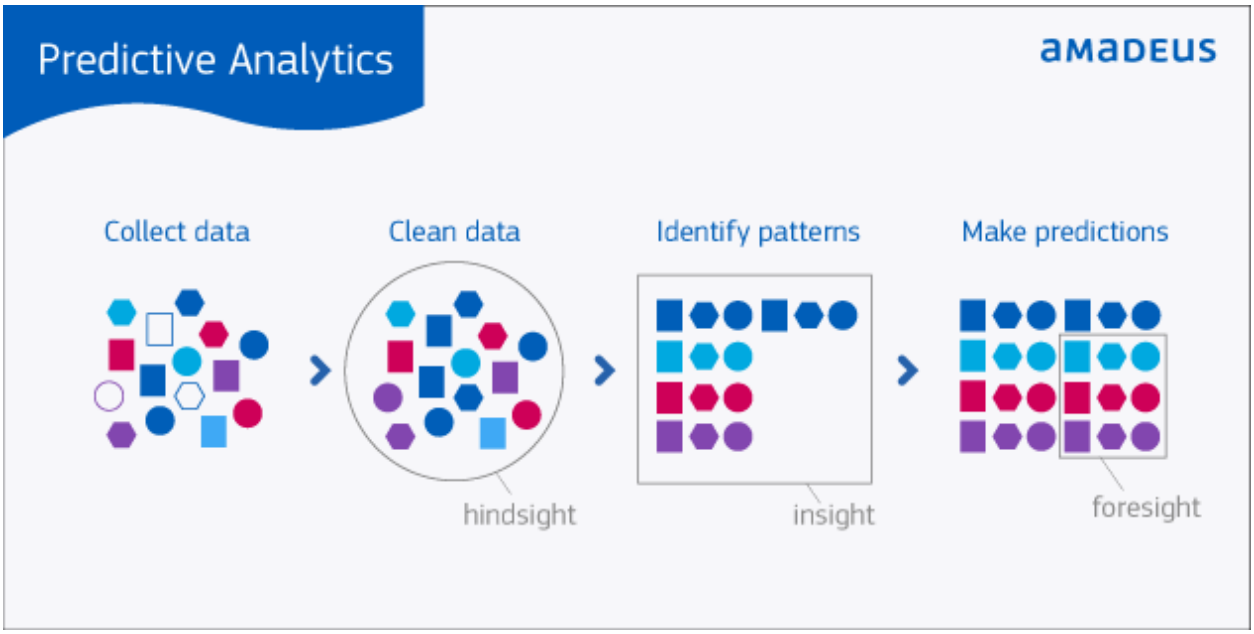
Predictive Analysis of Shares

Name: Swarup Mishal

NUID: 001620100

What is Predictive Analysis of Shares?

Predictive Analysis of Shares is a term that refers to the evaluation of a particular trading instrument, an investment sector or the market as a whole. Stock analysts attempt to determine the future activity of an instrument, sector or market. There are two basic types of stock analysis: fundamental analysis and technical analysis. Fundamental analysis concentrates on data from sources including financial records, economic reports, company assets and market share. Technical analysis focuses on the study of past market action to predict future price movement.



The project focuses to use available data in order to predict and increase profitability of the clients. The Company maintains stock market data from various data sources in its own database which will collectively be high in volume. Various types of data sources are,

- Investment management industry portfolio databases
- Custodian banks, asset servicers and TPAs
- Time Series Transactional & Government Data
- Investment Research Documents
- Satellite Imagery
- Social Media
- Events Data

- Marketing & Publishers
- Consumer Transaction Data
- Location Data and Consumer Behaviors

The data is generated with high velocity and variety each second. This data would be used to analyze the dependability of various companies and their shares. This analysis would then help the decision making process while investing Client's money in different shares. This generally includes buying and selling of investments within a portfolio. Investment management can also include banking and budgeting duties, as well as taxes. The term most often refers to portfolio management and the trading of securities to achieve a specific investment objective.

What role Big-Data plays in this project?

Big data is a catchphrase for a new way of conducting analysis. Big data principles are being adopted across many industries and in many varieties. However, adoption so far by investment managers has been limited. This may be creating a window of opportunity in the industry.

Investment managers who are able to harness this new approach could potentially create an information edge in their trading that would put them significantly ahead of their peers and allow them to profit from an "information arbitrage" between their expanded models and those of investment managers following more traditional analytic techniques.

Big data technology increases

- the volume of data that can be incorporated into investment models and
- the velocity at which that data can be processed.

Big data technologies rely on file-based databases in addition to traditional relational databases. As such, they can store not only structured data, but unstructured data as well. This means that new data sets can be added to a quantitative or systematic model more easily, without a lengthy cleansing, normalization, mapping and upload process. Thus with help of big data, we can improve and speed up the prediction process.

What strategy is required and why?

This project will require a build strategy for various reasons mentioned:

- **Off-the-shelf software cannot meet every need**: This project has specialized needs; custom software may be better qualified to meet them.
- **Canned solutions are rigid**: The vast majority of off-the-shelf software will not allow you to modify its functionality in a meaningful way. It may be difficult to add or subtract built-in features, leading to either too many or too few functions for your company.
- **Off-the-shelf software may not be compatible with other programs**: If you build your own software, you can integrate with a wider set of APIs from different software and data partners.

- **No restriction on budget:** Though the costs that are associated with building custom software may be one factor to be considered, but if the company has got sufficient funds, the return on investment can be well worth it.
- **Available technical proficiency**: If the company does have strong enough software team with the necessary skills to build out this custom software, one can create a custom based software easily.
- **Increased productivity**: Programs that are specifically designed with your needs in mind can enable your team to work faster and smarter. You can create one comprehensive technology platform as opposed to using multiple different programs. An integrated platform can yield major efficiency gains since all the data is one place and users do not have to switch between different websites as part of their workflow.
- **Competitive advantage**. When you rely on the same off-the-shelf software as your rival does, it is that much more difficult to outperform them. By designing your own technology that is ideally suited for your specific business operations, you can garner a competitive advantage relative to your competitors. That advantage grows as you invest more heavily in your proprietary systems.
- **Faster reaction time**. To build great custom software, you must first hire a stellar software development team. Once that team is in place, they can build a variety of products, tools, and systems. As your business needs change and as your industry evolves, being able to quickly shift technology strategies can mean the difference between market dominance and obsolesces.

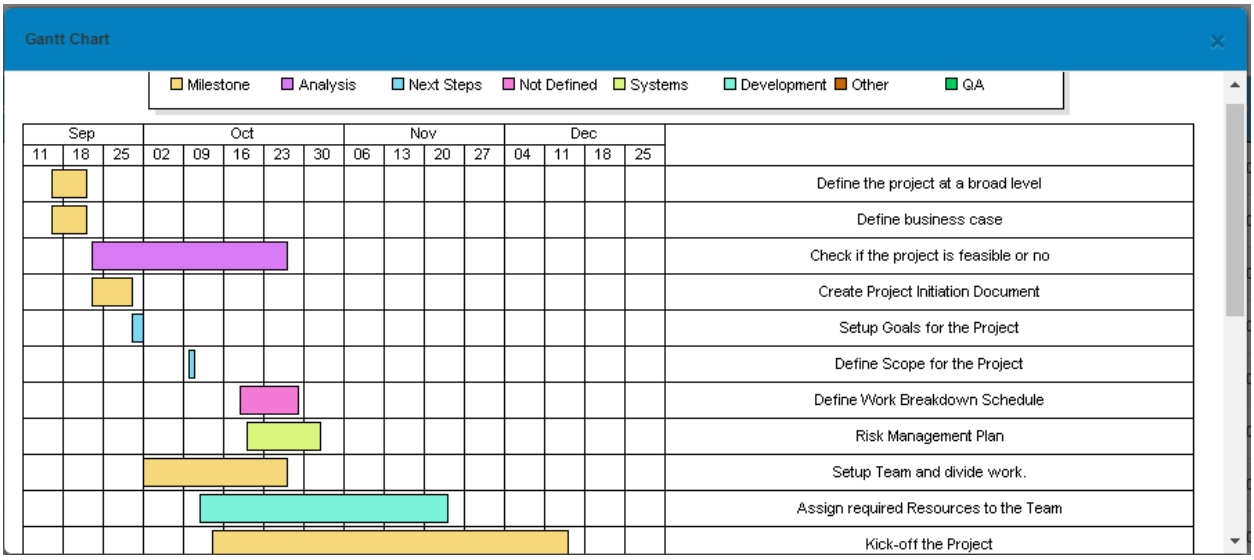
Strategy implemented for the Project

- **Problem:** The existing system has got some downside. The prediction process based on existing system is time consuming and gives less accurate results.
- **Solution**: The purpose of the Project is to improve the prediction process so that the it will help clients gain more profits from shares and this will ultimately lead to more clients approaching the company for consulting purposes.
- **Strategy:** The Company maintains stock market data from various sources in its own database which will collectively be high in volume. The data is generated with high velocity and variety each second. This data would be used to analyze the dependability of various companies and their shares. This analysis would then help the decision making process while investing Client's money in different shares.

	TRADITIONAL DATA ANALYSIS	BIG DATA ANALYSIS
DATA CAPTURE	Sampling (Randomly Selected Subset)	Mass Collection *N*=All
DATA QUALITY	Pristine & Curated	Raw & Varied
ANALYTIC GOAL	Extrapolate a Finding to Describe a Broad Population	Identify Correlations to Predict Outcomes

Roadmap for the Project

- **Project goals and objectives**: The objective project is to improve the prediction process by 30 % so that more clients would consult company while investing their money in various shares. As the number of clients' increase, this would fetch more profit for the company. The project would speed up the prediction process by 50%. As the prediction process will boost, it will indirectly save time.
- **A timeline indicating the schedule**: Basically timeline displays amount of time taken for each task during the entire project lifecycle. The entire project takes about 4 months to complete.



- **Important milestones and deliverables**: Important milestones for this project are,
    - Define the project at a broad level
    - Define business case
    - Create Project Initiation Document
    - Define Scope for the Project
    - Define Work Breakdown Schedule
    - Risk Management Plan
    - Setup Team and divide work.
    - Assign required Resources to the Team
    - Kick-off the Project
    - Project Manager directs and Manages Project Execution
    - If Needed modify Project Plan
    - Check if the Project is meeting Project Objectives
    - Check if the Big Data Solution is helping improve the prediction process.
    - Once the Project is complete, announce the Completion of Project successfully.
    - Evaluate Project success and failures
    - Prepare and save final Project Budget and Project Report for future.
- **Possible risks**: There are various risks related to this project.
    - If a resource leaves the company during the project duration it will lead to increase in the project duration. Also if the data from various sources is not up-to-date, it will hinder the prediction process.

    - Revenue is directly linked to market valuations, so a major fall in asset prices can cause a precipitous decline in revenues relative to costs.
    - Deviation from an expected outcome.
    - Above-average fund performance is difficult to sustain, and clients may not be patient during times of poor performance.
    - Successful fund managers are expensive and may be headhunted by competitors.
- **Dependencies**: Dependencies are the relationships of the preceding tasks to the succeeding tasks. Tasks may have multiple preceding tasks and multiple succeeding tasks. The most common dependency relationship is a finish-to-start relationship.
    - There would be a dependency integrating data from various data sources.
    - Another dependency would be between all the architectural components of the project.
    - Maintaining communication with the clients can also be considered as one of the dependency.

Project Characteristics

- **Consistency**: Project will be consistent. Regardless of the type of project you're trying to deliver or the type of client that you're working for, every project needs to be held to identical standards regarding what has to happen for it to be considered a success. This will allow you to identify certain areas of your process that may be redundant and can be eliminated, as well as avoid the type of bureaucratic practices that have derailed even the most well-laid plans.
- **Flexibility**: Project will be flexible. Flexibility means that you need to acknowledge that each specific project will have its own management needs, so you need to be ready to adjust your own way of doing things accordingly. More than that, certain types of issues will always arise unexpectedly, regardless of how detailed your plan is. Without that certain degree of flexibility, you don't have a hope of surviving the unexpected when it does rear its ugly head.
- **Transparency**: Project will be transparent. A successful project manager will know how to best schedule a project to guarantee its completion, for example. They will also know how to get information to the right people and help to distribute all important project dollars where they will have maximum impact. This all falls under the umbrella of transparency that should be practiced at all times.

Considerations regarding Data Governance

Almost any organization will say they make decisions based on data, or that they want to. Scratch the surface with a few questions about the quality of that data and confidence vanishes. Move on to discuss data security and see how quickly the subject changes. Data is clearly a key element in any business decision process today. As self-service is becoming the norm in data and analytics, more business users are demanding access to data to gain their own insights and drive localized initiatives.

The vast majorities of people have smart mobile devices and expect to access data from wherever they are. This any-device-anywhere culture will continue to grow as technology continues to become more powerful, and as data transfer rates improve globally. Let’s consider the risk of data in

transit and hatch a plan for it. Having a data implementation blueprint with executive support is imperative, it also needs to bring data analytics directly to the people who understand what questions to ask with regard to the insights it can provide. Modern data governance requires a true balancing act between enabling self-service analysis, and protecting sensitive business information. Most countries mandate data protection legislation for information about individual citizens. Additional layers of governance are also applied to public company financial information as well as other protections in healthcare and education.

Organizations don't just have to comply with these layers of governance, but proof of compliance is generally required. Data acquisition continues to build momentum as a key activity for all types of organizations, and with a myriad of data storage methods made available over the years, any longstanding company will likely still have legacy systems in place. Even with newer, on premise of cloud based solutions implemented, these legacy storage systems just never seem to fade away.

Finding a safe and effective management system for all the various data sources, including the random spreadsheets, coupled with a way to bring organizations to a higher standard of data quality, is a task that requires a combining of efforts for an IT and business approach. More precise data insights will show in terms of both increased revenues and reduced costs of operation ultimately driving better business outcomes. Organizations that understand their data can and will make better, more informed and faster decisions.

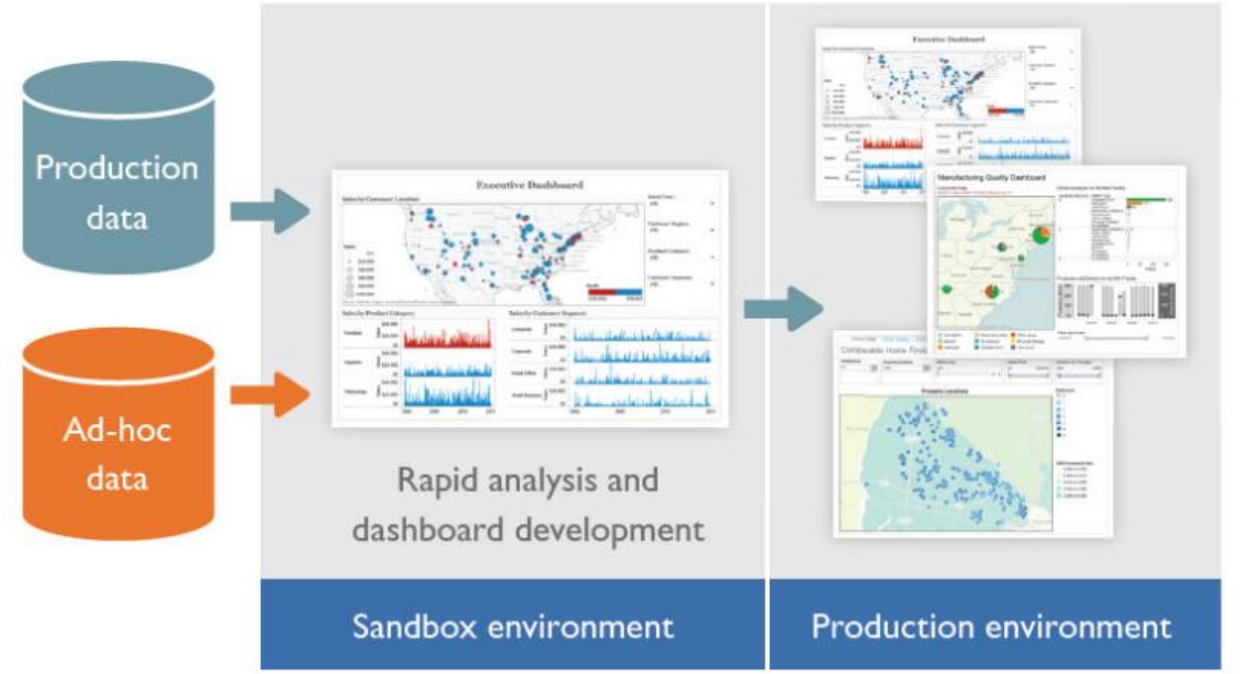
Three Essential Elements to be considered:

1) **The Team**:
    a. Started with a small project team to review existing data analytics use, and define a process to formalize and implement a wider self-service data analytics methodology.
    b. Included IT and business leaders on the team to give a balanced view of data use, governance and self-service needs.
    c. Involved at least one executive stakeholder to champion the data governance and utilization project.
2) **Data Quality:**
    a. Conducted a data audit to discover all the data sources in use, how current they are, and how many variants and versions of the same data are in existence and in use.
    b. Built a data infrastructure that allows for consolidation and guides users to shared data that is current and accurate.
    c. Defined a process for onboarding new data sources to make sure they meet quality standards and availability criteria. During this process, business users will usually identify and suggest new valuable sources of data, and IT will manage operationalizing it.
3) **Data Security and Compliance:**
    a. Made sure effective data classifications are in place so that data that requires protection for legal or regulatory reasons meets those requirements in an auditable way.
    b. Set levels of data access through leveraging your existing network access system, single sign-on through Active Directory, SAML or another solution is much easier to manage than having a separate data access protocol.
    c. Implemented tools that allow for different levels of access based on user privileges; this protects sensitive data while allowing access to higher-level analytics.

d. Included mobility and BYOD (bring your own data) into your plans as these are one of the largest sources of data leakage. 4. Report and Dashboard Governance
e. Provided self-service analytics with clear guidelines on how to access the right data
f. Implemented a process to review and approve reports and dashboards to production, while providing a “sandbox” for users to develop and test reports.
Communicated the full process and appoint business and IT data stakeholders as points of contact and advocates.

Report and Dashboard Governance should also be considered:

a. Provide self-service analytics with clear guidelines on how to access the right data.
b. Implement a process to review and approve reports and dashboards to production, while providing a “sandbox” for users to develop and test reports.
c. Communicate the full process and appoint business and IT data stakeholders as points of contact and advocates.



Conclusions:

The key to data governance in an age of self-service analytics is to make IT a partner to the business. Data governance is not about locking everything down to a few privileged users, but about enabling broad groups of users with the appropriate controls. Business users get secure access to shared production data sources managed by IT, so they don’t have to go fishing for the right data or worry whether data is secure. IT has the opportunity to set security controls and business practices, helping them keep their data accurate and secure. When done right, data governance is about providing the right data to the right people whenever they need it, where ever they need it, so they can answer their own questions.

Functional requirements

We will have to take following functional requirements into considerations:

- **Historical Data**: Developing an effective big data project heavily revolves around the ability to handle historical data. Most of the times, historical data is present in the form of relational tables that need to be migrated to an unstructured format in order to accommodate big data methodologies.
- **Authentication and Authorization**: Authentication and Authorization Requirements play an important role while defining functional requirements for system. They provide user access information and provide security to the system.
    - Authentication: It involves management of system access with respect to user. Authentication is a process by which you verify that someone is who they claim they are.
    - Authorization: Authorization is the process of establishing if the user (who is already authenticated), is permitted to have access to a resource. Authorization determines what a user is and is not allowed to do.
- **Audit Tracking**: Audit tracking is one of the functional requirement for a system which is of significant importance. An audit trail or audit tracking (also called audit log) is a security-relevant chronological record, set of records, and/or destination and source of records that provide documentary evidence of the sequence of activities that have affected at any time a specific operation, procedure, or event. Audit trails are used to record customer activity in e-commerce. A company might also use an audit trail to provide a basis for account reconciliation, to provide a historical report to plan and support budgets, and to provide a record of sales in case of a tax audit.
- **BI and Reporting**: Integrate the selected BI tool with the database to pull in data. Creation of BI use cases. Create and design dashboards. Visualize the data. Setup reporting for management.
- **Data Integration and Storage**: Model and design of database (Conceptual, Logical, Physical model). Identify the data source and how it has to be captured. Establish connections with the data sources. Data staging. Data profiling. Setup ETL (cleansing, filtering, joins etc.) jobs to extract and load data to the designed database. Getting the data in target databases for further analysis.
- **Business Rules, Administrative functions**: The logical layers of a big data solution help define and categorize the various components required for a big data solution that must address the functional requirements. This set of logical layers outlines the critical components of a big data solution from the point where data is acquired from various data sources to the analysis required to derive business insight to the processes, devices, and humans who need the insight.

Non-functional requirements

We will have to take following non-functional requirements into consideration:

- **Scaling with Growing Data Volume and Workload**: The system shall be scalable so that the increasing data volumes can be processed and analyzed as per the requirement. It is expected of

the system that by adding hardware, the system would become more scalable i.e., linearly scalable by a factor of at least <define a scaling factor> as defined by the project data volumes.

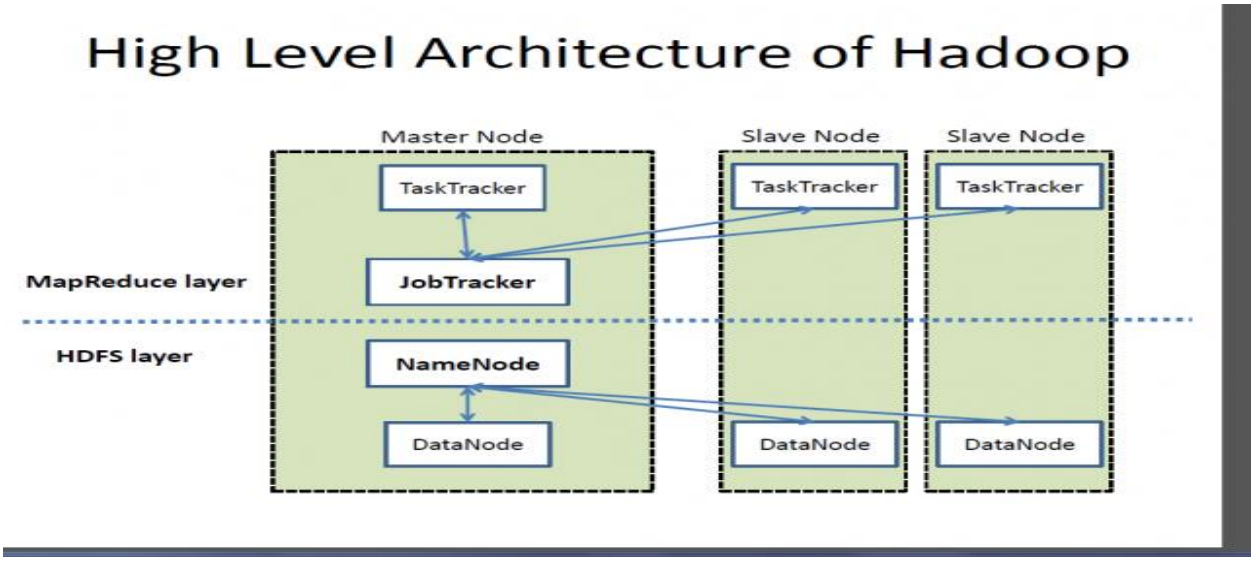
- **Handling Streaming Data while it is flowing in**: The system shall handle and is capable of analyzing the data that is flowing from a variety of sources up an inflow rate of <define inflow rate>
- **Creating Insights from Streaming Data – Timeliness**: The system should be capable of creating insights from the streaming data into the system within an acceptable time frame of <define time> while the data is flow at a rate defined earlier.
- **Acquiring and Processing Streaming Data - Acquisition Rate**: The system is expected to acquire, pre-process and store the data which is streaming into the system at the specified rate and with an acquisition rate <define acquisition rate>
- **Handling Data in Multiple Structures and Formats**: The system is expected to handle the data coming from multiple inputs like logs, data warehouses etc., so it is expected of the system to handle various formats of data as <define all the data formats>. Also the sources of data have to be specified <name all the sources>. All the data that is acquired from all the sources is expected to be analyzed and stored so that the performance can be as specified earlier.

Architectural components and reasons for choosing each

Various components have been used for implementing big data technologies. Each component and the reason for selecting it has been explained in detail below.

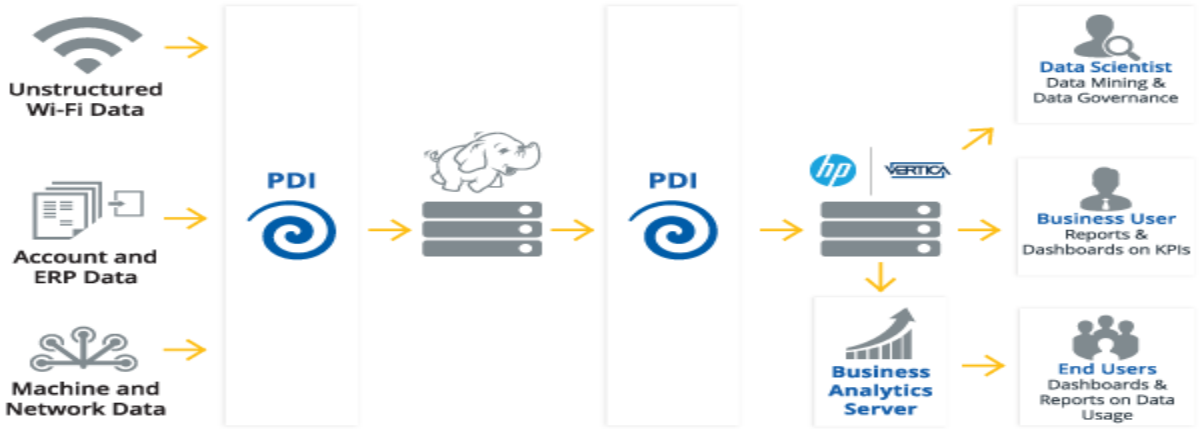
- **Framework**: Hadoop
    - Hadoop is the Open Source database technology that has driven much of the file-based technology adoption currently under way. What makes Hadoop different from previous technologies is its inherent design for large, unstructured data sets within distributed hardware and application environments. Also, the Hadoop stack has matured over the years to provide a set of tools and data processing layers to meet some of the business challenges relevant to investment manager needs.
    - The database is comprised of two major components – HDFS, the file store component, and MapReduce, a job scheduler with other functions that manages processing of data requests. HDFS stores and decomposes files into smaller blocks of information over multiple servers and disks and constantly checks these servers for disk problems. This inherently makes HDFS scalable and fault tolerant, because stored data is replicated to another server or storage node if there is an issue.
    - Also, during processing, HDFS load-balances and optimizes processing across multiple stores of the same data. In the simplest explanation, MapReduce enables jobs to be scheduled against stored data in nodes and clusters and split into smaller blocks and even smaller individual field level records. Jobs are also mapped to where the data is stored.
    - Newer generations of Hadoop decouple MapReduce's resource management and scheduling capabilities from the data processing component, enabling Hadoop to support more varied processing approaches and a broader array of applications. For

example, Hadoop clusters can now run interactive querying and streaming data applications simultaneously with MapReduce batch jobs.

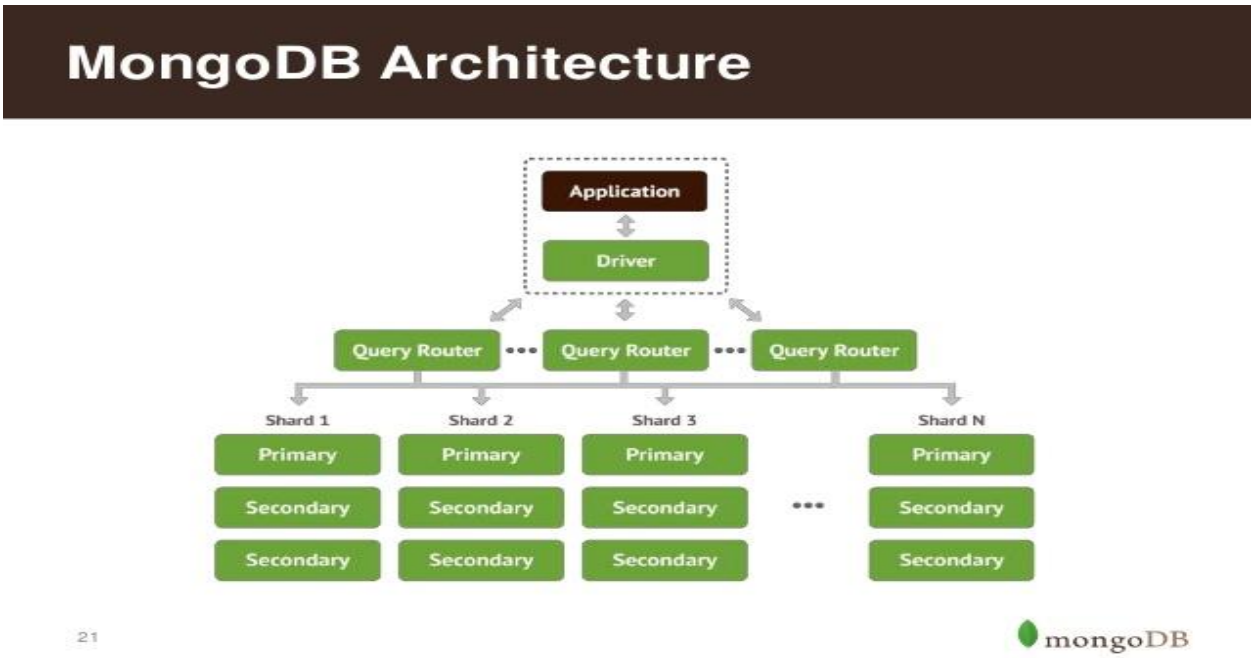


- **Integration**: Pentaho
  - Hadoop data integration presents IT organizations with challenges, including acquiring new technology skillsets, finding the right developers, and effectively linking Hadoop with existing operational systems and data warehouses.
  - Pentaho's intuitive and powerful platform is built to tackle these challenges head-on, but delivering accelerated productivity and time to value is just the beginning. Pentaho helps teams manage complex data transformations and enables them to operationalize Hadoop and Spark as part of an end-to-end data pipeline, ensuring the delivery of governed analytics.
  - Intuitive visual interface to integrate and blend Hadoop data with virtually any other source – including relational databases, NoSQL stores, enterprise applications, and more
  - Ability to design MapReduce jobs 15 times faster than hand-coding approaches
  - Native MapReduce integration that executes complex transformation and blending logic in-cluster, while scaling linearly with Hadoop
  - Deep integration with the Hadoop ecosystem including SQL on Spark connectivity, the ability to orchestrate a variety of Spark applications, and compatibility with Kafka, YARN, Oozie, Sqoop, and more
  - Automation to rapidly accelerate the ingestion and onboarding of hundreds or thousands of diverse and changing data sources into Hadoop
  - Support for leading Hadoop distributions, including Cloudera, Hortonworks, Amazon EMR, and MapR, with maximum portability of jobs and transformations between Hadoop platforms
  - Solution approach to deliver on-demand data sets from Hadoop, including governed self-service analytics for large production user bases
  - Full array of visualizations, reports, and ad hoc analysis, including connectivity to Hive, Impala, and analytic databases such as Vertica and Redshift

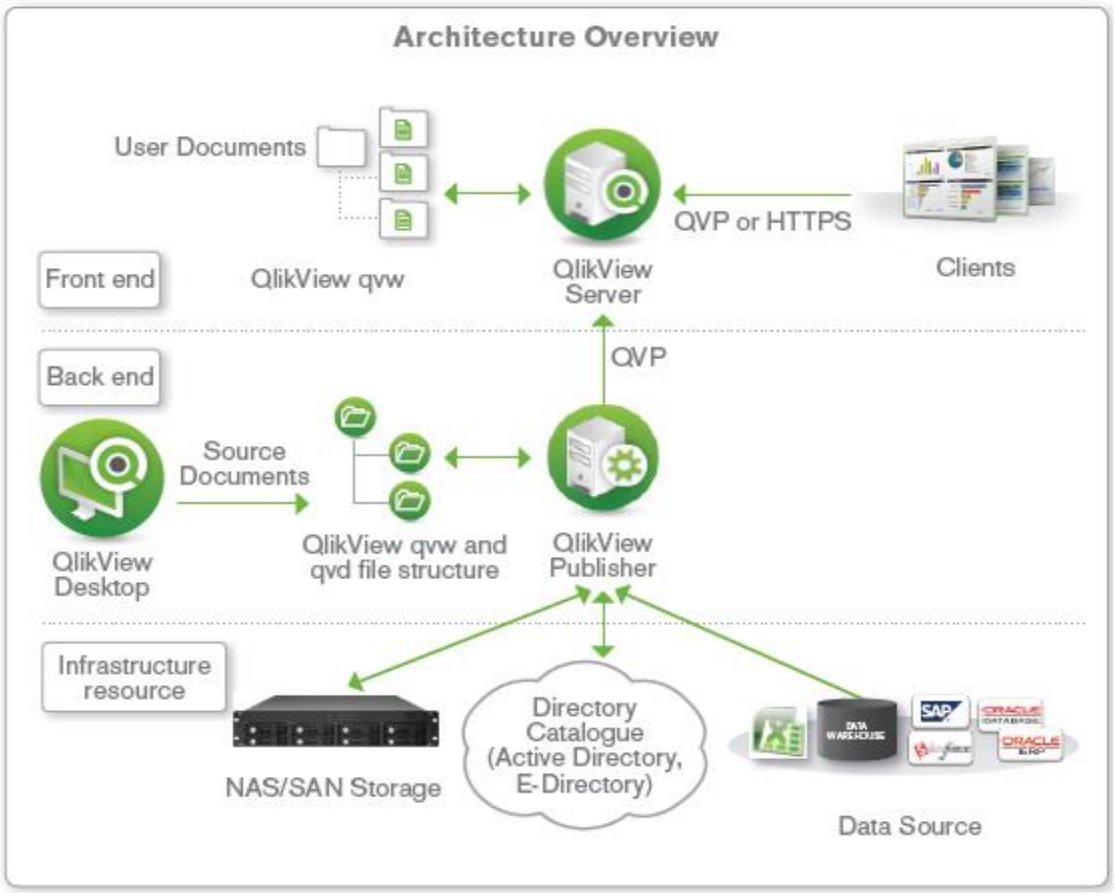
- Analytics that can be seamlessly embedded into crucial business applications to drive data monetization with customers and partners
- Ability to incorporate predictive models from R and Weka into the data flow, driving actionable results while minimizing the data prep burden

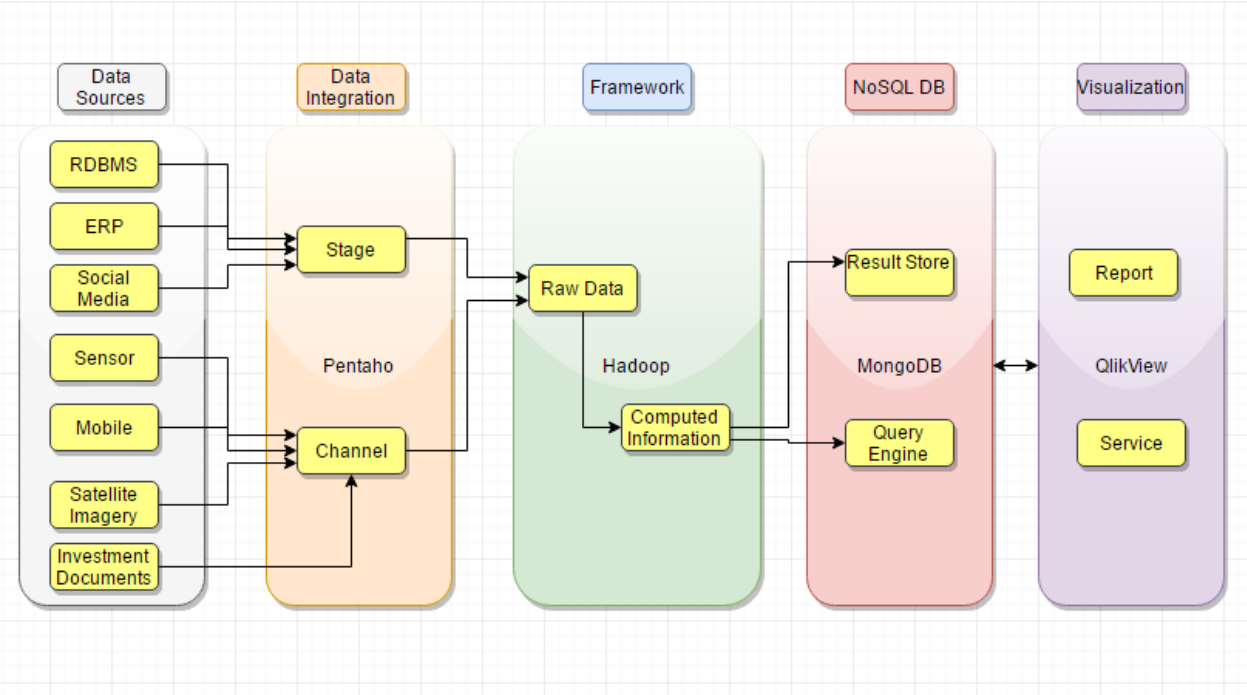


- **Language**: Java
  - Java's code is portable and platform independent which is based on Write Once Run Anywhere.
  - Java is fairly simple language to learn to develop a complex and large project. Java project are easy to manage.
  - Java has a huge collection of open source libraries than any other language.
  - Java programs crashes less catastrophically and it is easier to debug an issue in Java then C or C++.
  - Garbage Collection techniques is easier to reuse or extend to a project need in Java.
- **Database**: MongoDB
  - MongoDB is a document-oriented database or is described as a No-SQL database. Founded more as a database to operate large websites, Mongo is one of the back-ends for large web businesses, such as eBay, Foursquare and the NY Times.
  - The focus of Mongo is more operational and it is designed to be distributed, easy to program and perform read-write functions in a scalable, redundant way. It uses JavaScript Object Notation 'JSON-like' documents with dynamic schemas (MongoDB calls the format BSON), making the integration of data in certain types of applications easier and faster.
  - MongoDB retains some of the characteristics of RMDBs, making it easier to query and port relational database data to MongoDB. Documents can be indexed and queried and secondary indexing is also supported. MongoDB works in conjunction with Hadoop on more analytic, batch-focused use cases.
  - Algorithmic trading, where high-speed and high throughput are required, is one use case for which MongoDB is not well suited. Such a high throughput data processing use case is more the domain of established in-memory, column databases like kdb+.



- **Visualization**: QlikView
    - In QlikView, one single application provides access to various types of data, allowing Management to easily come to new insights.
    - QlikView makes it possible to perform an infinite number of analyses. Every angle of approach can be viewed and analysed.
    - Due to its social and real-time business discovery options, QlikView enables teams to take collaborate decisions.
    - QlikView makes it possible to interpret data quickly with visual and dynamic dashboards, apps and statistics.
    - QlikView is true self-service Business Intelligence, empowering departments to get to work more quickly and with less dependence on IT.
    - QlikView combines various data sources and handles reporting and data transfers so that IT professionals can focus on their core tasks.
    - QlikView is simple to implement and requires little maintenance time.
    - QlikView is constructed on in-memory technology which maintains the relationships between the various data.
    - QlikView complies with extremely strict security standards so that business-critical data is always protected.



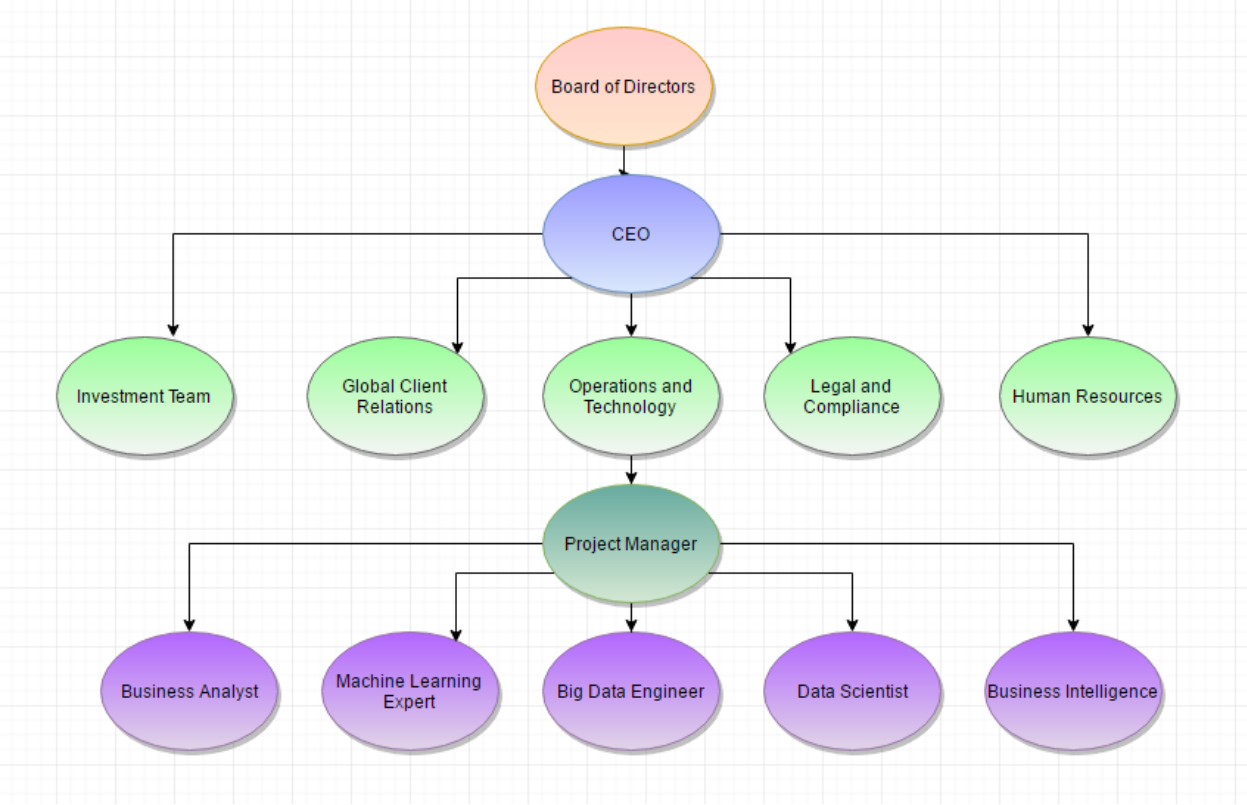


Overall Architecture in a nutshell.

Organization Structure

Let's see the broad picture of the organization.

a) **Board of Directors**: The board has the duties of making sure customers' deposits are secured and invested safely, interest is paid to depositors, and that the customers' principal is available to them on request.
b) **Chief Executive Officer**: A chief executive officer (CEO) is the highest-ranking executive in a company, and their primary responsibilities include making major corporate decisions, managing the overall operations and resources of a company, and acting as the main point of communication between the board of directors and corporate operations. Following teams fall under the CEO.
    i) **Investment Teams**: They deliver high-quality strategic advice and creative financing solutions to clients, including mergers, acquisitions, financing, and risk management transactions.
    ii) **Global Client Relations**: Their role is to manage and protect the relationship between the company and its most important clients globally. Client relationship managers aim to maximize long-term revenue opportunities by becoming trusted advisers to clients and strengthening their loyalty.
    iii) **Legal and Compliance**: The goal of the Legal and Compliance department is to provide quality legal services in-house by professionals who are intimately familiar with the mission, activities, and business operations and opportunities.
    iv) **Human Resources**: The role of human resources is to plan, develop, and administer policies and programs designed to make expeditious use of an organization's human resources.
    v) **Operations and Technology**: They design, manufacture, distribute, and deliver a product or a service. In essence, operations connect the dots across the value chain. This is where our team comes into picture under Information Technology.
        (1) Project Manager: First of all, we will need a Project Manager to keep an eye on and manage entire team.
        (2) Business analyst: They have intimate knowledge of your industry and company, and they analyze business-level data to produce actionable insights.
        (3) Machine-learning expert: This person is statistically minded, with experience in programming and building data models. This person develops algorithms and crunches numbers in order to help answer questions and make predictions.
        (4) Big-Data Engineer Team: The data engineer is concerned with the capture, storage, and processing of the data itself.
        (5) Data Scientist Team: Ideally, this person has domain knowledge, a statistical analysis background, and basic understanding of computer science in order to extract information from the data to answer business questions.
        (6) Business Intelligence Team: Their main function is to support a company's decision-making process and to help knowledge workers, such as managers and research analysts, make better and faster decisions.



Organization Structure in a nutshell.

Scalability

Since we are using MongoDB we have,

- Cluster Scale - Distributing the database across 100+ nodes, often in multiple data centers.
- Performance Scale - Sustaining 100,000+ database read and writes per second while maintaining strict latency SLAs.
- Data Scale - Storing 1 billion+ documents in the database.

With the help of Pentaho, we are able to scale,

- By tightly coupling a high-performance business intelligence with data integration in a single platform, Pentaho Business Analytics provides a scalable solution that can address enterprise requirements in organizations of all sizes.
- When attempting to process the vast amounts of data collected on a daily basis, it is critical to have a Data Integration solution that is not only easy to use but easily extendable.
- Pentaho Data Integration achieves this extensibility with its open architecture, component stack and object library which can be used to build a scalable and highly available ETL solution without exhaustive training and no code to write, compile or maintain.

Hadoop helps in scalability as,

- Horizontal-scaling it is often easier to scale dynamically by adding more machines into the existing pool.
- Linear / Horizontal scalability: more nodes can do more work within the same time
- Linear on data size
- Linear on compute resources

Manageability

Company will take responsibility for managing this entire Project.

MongoDB is managed by Mongoclient.

- Mongoclient is a completely free and open-source mongodb management tool. It's written in MeteorJS . Additionally, it's fully responsive and have a nice look and feel. Available on most platforms including Mac, Linux, Windows with portable distributions, as an advantage of responsive design and MeteorJS, it's easier to use Mongoclient on most mobile platforms.
- MongoDB Compass is built by the team that engineers the database, MongoDB Compass is a GUI for MongoDB.
- MongoDB Atlas is a hosted database as a service from the team that engineers the database.
- MongoDB Cloud Manager is a management platform for MongoDB, delivered as a cloud service.

Pentaho is managed as,

- Pentaho Data Integration provides a declarative approach to ETL where you specify what to do rather than how to do it. It includes a transformation library with over 70 mapping objects. It includes data warehousing capability for slowly changing and junk Dimensions. Includes support for multiple data sources including over 25 open source and proprietary database platforms, flat files, Excel documents, and more. The architecture is extensible with a plug-in mechanism.

Managing big data can be daunting, so one should seek a distribution for Hadoop that lowers the administrative burden. Manageability, including multi-tenancy, is key to ensuring production success and continued growth.

Security

MongoDB is secured using,

- Authorization Role-Based Access Control Enable Auth Manage Users and Roles
- Transport Encryption Configure mongod and mongos for TLS/SSL TLS/SSL Configuration for Clients
- Enterprise Security support Kerberos Authentication LDAP Proxy Authority Authentication Encryption at Rest Auditing

Pentaho security can be maintained using valid user authentication and authorisation.

Security features of Hadoop consist of authentication, service level authorization, authentication for Web consoles and data confidentiality.

- Authentication - By default Hadoop runs in non-secure mode in which no actual authentication is required. By configuring, Hadoop runs in secure mode, each user and service needs to be authenticated by Kerberos (a computer network authentication protocol) in order to use Hadoop services.
- Service Level Authorization - Service Level Authorization: It is the initial authorization mechanism to ensure clients connecting to a particular Hadoop service have the necessary, preconfigured, permissions and are authorized to access the given service.
- For example, a MapReduce cluster can use this mechanism to allow a configured list of users/groups to submit jobs.
- Authentication of Web Consoles - By default Hadoop HTTP web-consoles (JobTracker, NameNode, TaskTrackers and DataNodes) allow access without any form of authentication. Similarly, to Hadoop RPC, Hadoop HTTP web-consoles can be configured to require Kerberos authentication using HTTP SPNEGO protocol (supported by browsers like Firefox and Internet Explorer).
- In addition, Hadoop HTTP web-consoles support the equivalent of Hadoop's Pseudo/Simple authentication. If this option is enabled, user must specify their user name in the first browser interaction using the user.name query string parameter. Data Encryption on RPC
- The data transferred between hadoop services and clients.
- Setting hadoop.rpc.protection to "privacy" in the core-site.xml activate data encryption. Data Encryption on Block data transfer.
- You need to set dfs.encrypt.data.transfer to "true" in the hdfssite.xml in order to activate data encryption for data transfer protocol of DataNode. Data Encryption on HTTP.
- Data transfer between Web-console and clients are protected by using SSL(HTTPS).

References

http://www.simcorp.com/en/insights/journal/big-data-what-it-is-and-what-it-might-mean-to-investment-managers

https://www.managedfunds.org/wp-content/uploads/2016/02/CITI-_BigData_Final_Web.pdf

https://www.bnymellon.com/_global-assets/pdf/our-thinking/business-insights/big-data-and-investment-mangement.pdf

http://www.cmegroup.com/education/files/big-data-investment-management-the-potential-to-quantify-traditionally-qualitative-factors.pdf

http://www.tableau.com/learn/whitepapers/data-governance-self-service-analytics

https://www.rimes.com/insights/data-governance-becoming-vital-cog-investment/

https://opsdog.com/industries/asset-management/asset-management-organization-chart

http://www.projectinsight.net/project-management-basics/task-dependencies

Thank You!