# Big Data and Data Science

## 1 Introduction

Generate, store and analyse larger amounts of data:

- **Big Data:** Volume, variety and velocity it is produced
- **Data Scientists:** People who mine and develop predictive, machine learning and prescriptive models and analytics
- **Data Science:** Applies data mining, statistical analysis and machine learning with data integration and data modelling to predict behaviours. Forward Looking. BI describes past trends – rear-window view.

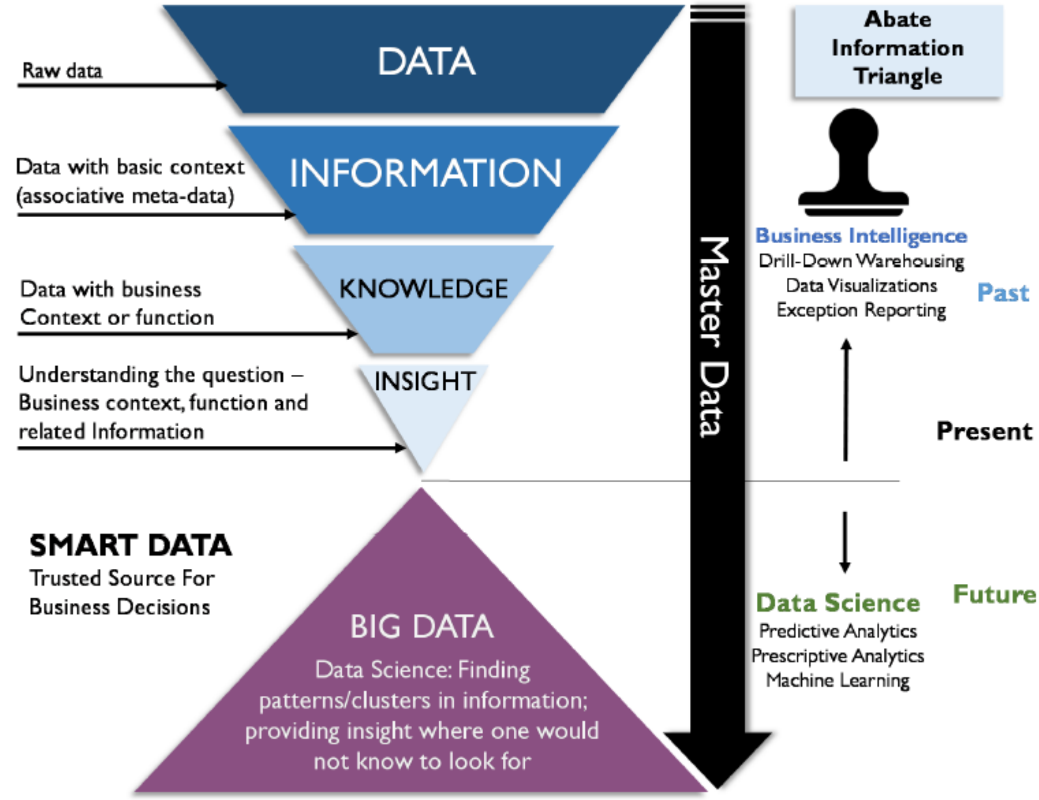
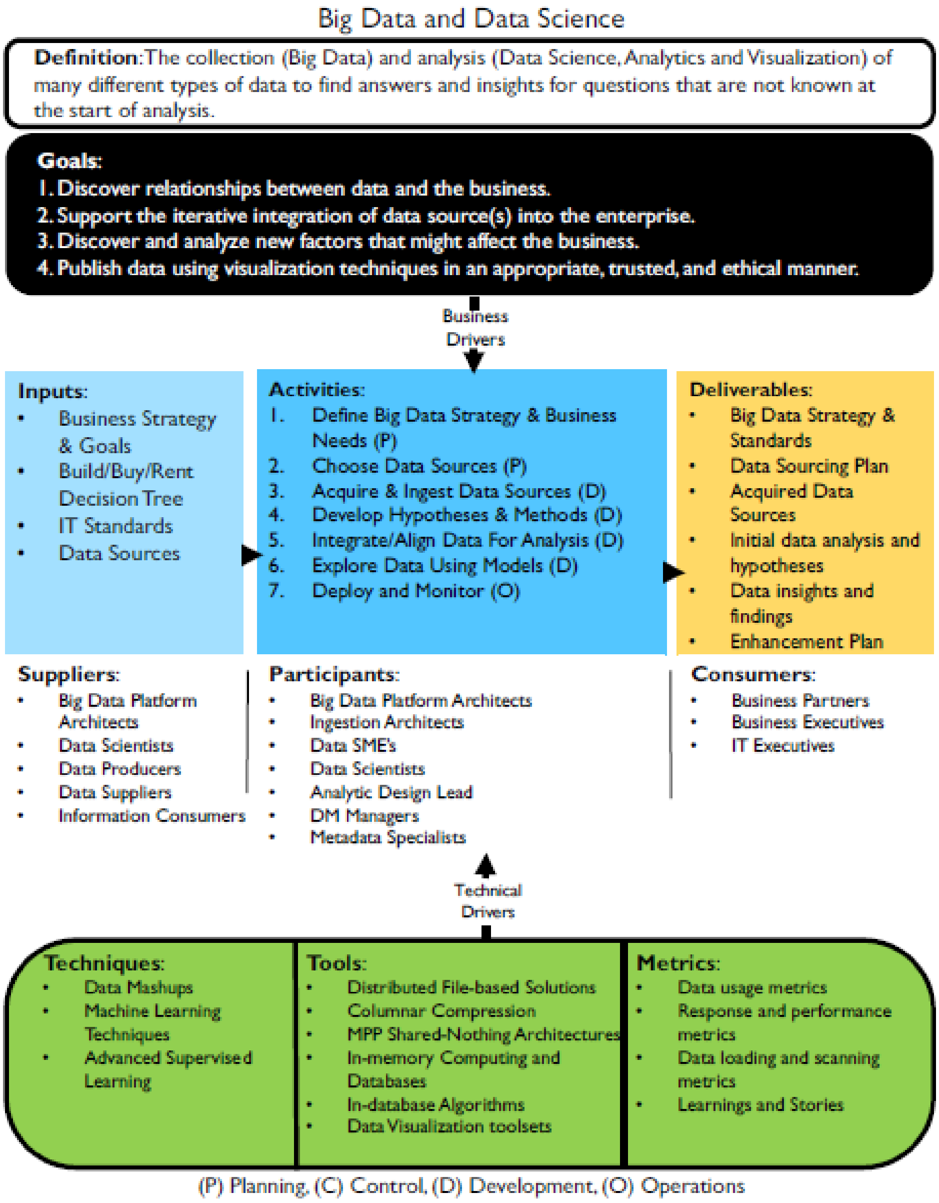


Figure 96 Abate Information Triangle

Different management and storage. Big Data relies on ELT – Loading and then transforming – and is not stored in the relational model. Speed and volume of data requires different approaches to Integration, Metadata and Data Quality management.

### 1.1 Business Drivers

- The desire to discover and act on business opportunities through applying techniques on diversely generated data.
- Data Science can improve operations
- Machine learning for automation of complex time-consuming activities



Big Data and Data Science

**Definition:** The collection (Big Data) and analysis (Data Science, Analytics and Visualization) of many different types of data to find answers and insights for questions that are not known at the start of analysis.

**Goals:**
1. Discover relationships between data and the business.
2. Support the iterative integration of data source(s) into the enterprise.
3. Discover and analyze new factors that might affect the business.
4. Publish data using visualization techniques in an appropriate, trusted, and ethical manner.

Business Drivers

**Inputs:**
- Business Strategy & Goals
- Build/Buy/Rent Decision Tree
- IT Standards
- Data Sources

**Activities:**
1. Define Big Data Strategy & Business Needs (P)
2. Choose Data Sources (P)
3. Acquire & Ingest Data Sources (D)
4. Develop Hypotheses & Methods (D)
5. Integrate/Align Data For Analysis (D)
6. Explore Data Using Models (D)
7. Deploy and Monitor (O)

**Deliverables:**
- Big Data Strategy & Standards
- Data Sourcing Plan
- Acquired Data Sources
- Initial data analysis and hypotheses
- Data insights and findings
- Enhancement Plan

**Suppliers:**
- Big Data Platform Architects
- Data Scientists
- Data Producers
- Data Suppliers
- Information Consumers

**Participants:**
- Big Data Platform Architects
- Ingestion Architects
- Data SME's
- Data Scientists
- Analytic Design Lead
- DM Managers
- Metadata Specialists

**Consumers:**
- Business Partners
- Business Executives
- IT Executives

Technical Drivers

**Techniques:**
- Data Mashups
- Machine Learning Techniques
- Advanced Supervised Learning

**Tools:**
- Distributed File-based Solutions
- Columnar Compression
- MPP Shared-Nothing Architectures
- In-memory Computing and Databases
- In-database Algorithms
- Data Visualization toolsets

**Metrics:**
- Data usage metrics
- Response and performance metrics
- Data loading and scanning metrics
- Learnings and Stories

(P) Planning, (C) Control, (D) Development, (O) Operations

Figure 97 Context Diagram: Big Data and Data Science

## 1.2 Principles

Important to manage Metadata related to Big Data sources for inventory, their origins and value.

## 1.3 Essential Concepts

### 1.3.1 Data Science

Developing predictive models that explore data content patterns. Based on hypothesis which may be statistically confirmed by historical data. Iterative inclusion of data sources into models that develop insights.

Data science depends on:

- **Rich data sources:** Potential to show otherwise invisible patterns in organisational or customer behaviour.
- **Information alignment and analysis:** Techniques to understand data content and combine data sets for meaningful patterns

- **Information delivery:** Visualisations of results of insight gained from mathematical models
- **Presentation of findings and data insights:** Presentation and sharing of insights

Table 32 Analytics Progression

| DW / Traditional BI | Data Science | |
|---|---|---|
| Descriptive | Predictive | Prescriptive |
| Hindsight | Insight | Foresight |
| Based on history:<br>What happened?<br>Why did it happen? | Based on predictive models:<br>What is likely to happen? | Based on scenarios:<br>What should we do to make things happen? |

### 1.3.2 The Data Science Process

Data Science follows the scientific method of refining knowledge by making observations and testing hypotheses, observing results and formulating theories to explain results.  The output of each phase in the diagram is input to the next.

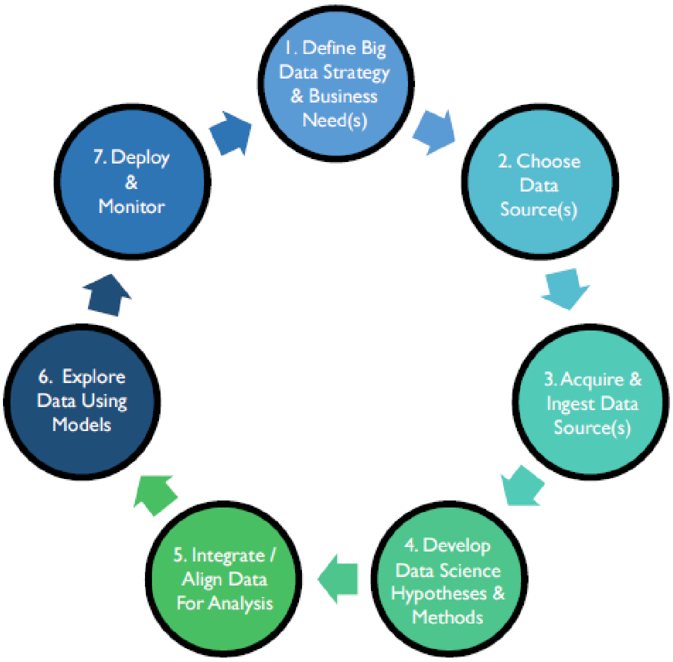

1. Define Big Data Strategy & Business Need(s)
2. Choose Data Source(s)
3. Acquire & Ingest Data Source(s)
4. Develop Data Science Hypotheses & Methods
5. Integrate / Align Data For Analysis
6. Explore Data Using Models
7. Deploy & Monitor

Figure 98 Data Science Process

- **Define Big Data strategy and business needs:** Requirements that identify desired outcomes with measurable benefits
- **Choose data sources:** Identify gaps in current data base and find sources
- **Acquire and ingest data stores:**
- **Develop Data Science hypotheses and methods:**
- **Integrate and align data for analysis:** Data integration and cleansing for quality (Wrangling and Munging)
- **Explore data using models:**
    - Apply statistical analysis and machine learning algorithms
    - Validate, train and evolve model
    - New hypotheses may be introduced
- **Deploy and monitor:** often becomes warehouse

### 1.3.3 Big Data

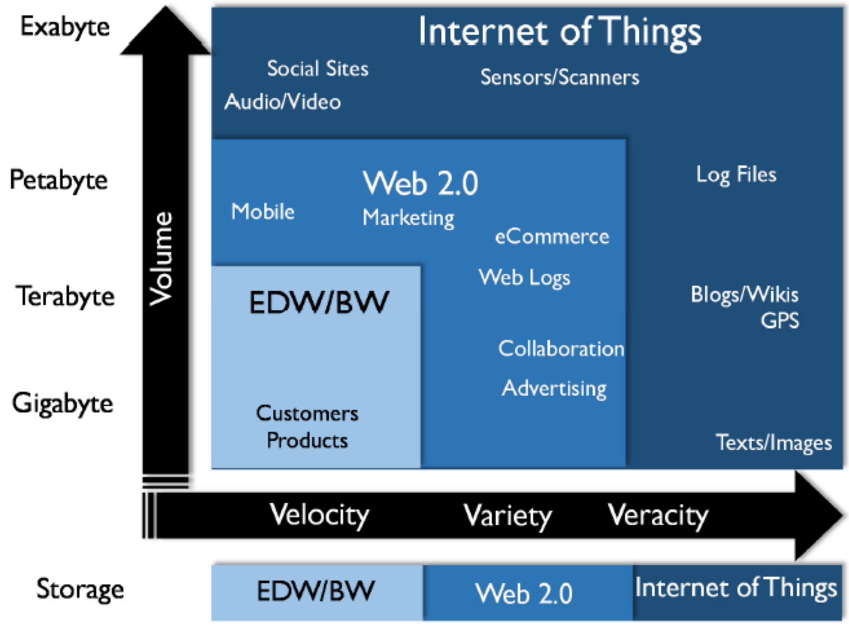


Figure 99 Data Storage Challenges[88]

- **Volume:** Amount – billions of records - >100 Terabytes
- **Velocity:** Speed at which data is captured, generated or shared – often analysed at real-time
- **Variety / Variability:** Forms data is captured or delivered
- **Viscosity:** How difficult the data is to use and integrate
- **Volatility:** How often the data changes
- **Veracity:** How trustworthy the data is

### 1.3.4 Big Data Architecture Components

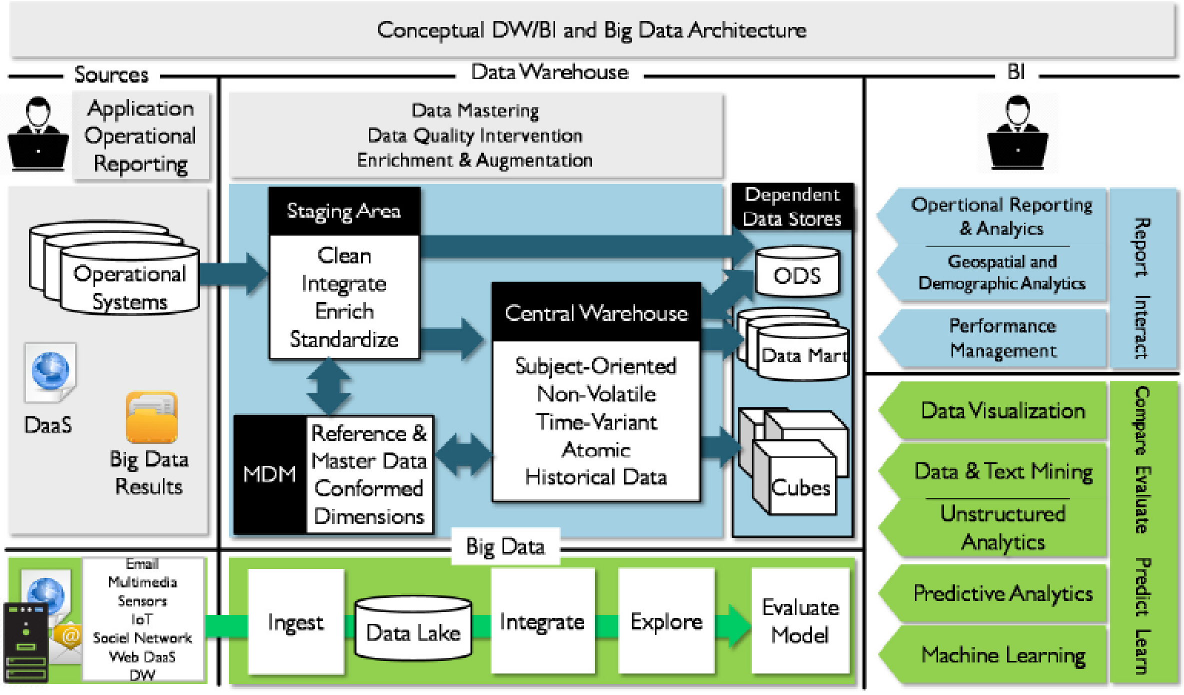


Figure 100 Conceptual DW/BI and Big Data Architecture

### 1.3.5 Sources of Big Data

Internet of Things, email, social media, online orders, video games, phones, POS devices, surveillance systems, sensors, medical monitoring, satellites, military etc.

### 1.3.6 Data Lake

An environment where a vast amount of data of various types and structures can be ingested, stored, assessed and analysed:

- Environment for Data Scientists to mine and analyse data
- Central storage for raw data with minimal transformation
- Alternate storage for detailed historical data warehouse data
- Online archive for records
- Environment to ingest streaming data with automated pattern identification

Manage Metadata to prevent it becoming a data swamp.

### 1.3.7 Services-Based Architecture

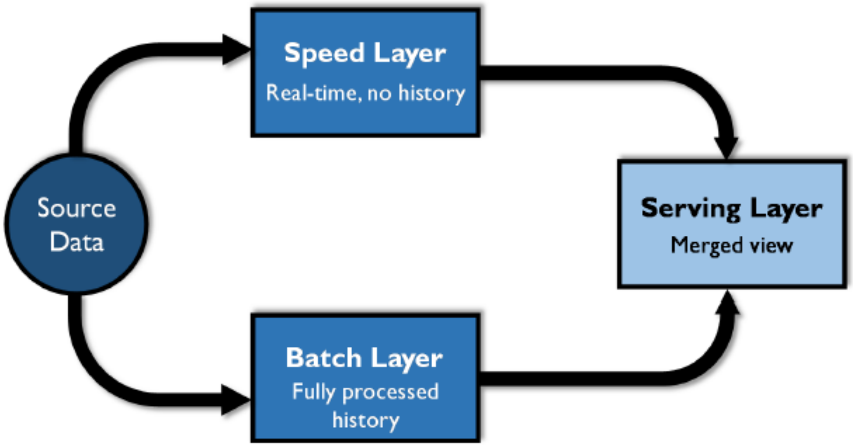


Figure 101 Services-based Architecture

Also called Lambda Architecture, referred to in Ch 5 p181, CAP / Brewers Theorem.

SBA provides immediate (maybe not complete or accurate) data. Updates complete, accurate historical data set using same source

SBA has 3 main components:

- **Batch layer:** A data lake, historical, structure-over-time component, every transaction is an *insert*.
- **Speed layer:** Only real-time data, Operational Data Store (ODS), all transactions are *updates* or inserts
- **Serving layer:** Interface to join batch and speed layers, uses Metadata to determine where to "serve" the data

Data is loaded into both Batch and Speed layers simultaneously, creating a current state, and a history layer.

### 1.3.8 Machine Learning

The construction and study of learning algorithms. Subfield of Artificial Intelligence. Three types:

- **Supervised learning:** Based on generalised rules e.g. separating SPAM emails

- **Unsupervised Learning:** Identifying hidden patterns i.e. data mining
- **Reinforcement learning:** Based on achieving a goal e.g. beating a chess opponent

Ethical implications:

- Transparency:
    - It is not clear how Deep Learning Neural Networks (DLNN) learn
    - Need to see how decisions are made

### 1.3.9 Sentiment Analysis

Looking for key words in semi-structured data using Natural Language Processing (NLP) to look for sentiment. Requires understanding of the meaning of a post.

### 1.3.10 Data and Text Mining

**Data Mining** is an offshoot of machine learning that reveals patterns in data using algorithms. Unsupervised learning where algorithms are applied to a data set without knowledge of outcomes, to reveal patterns and relationships.

**Text mining** analyses documents with text analysis and data mining techniques to classify content automatically into workflow guided and SME-directed ontologies.

Data and text mining techniques:

- **Profiling:** Characterise behaviour norms of an individual, group or population for anomaly detection (e.g. fraud)
- **Data reduction:** Make a similar smaller data set from a large one for easier analysis.
- **Association:** Unsupervised learning process to find relationships based on transactions
- **Clustering:** Group elements by shared characteristics
- **Self-organising maps:** Neural network method of cluster analysis by reducing dimensionality in the evaluation space. Also called Kohonen Maps or topologically ordered maps.

### 1.3.11 Predictive Analytics

Subfield of supervised learning. The development of probability models based on variables related to possible events. When it receives some information, the model triggers a response by the organisation. A forecast is a simple predictive model.

### 1.3.12 Prescriptive Analytics

Prescriptive analytics anticipates what will happen, when it will happen and implies why it will happen. Shows implications of various decisions, and can suggest how to take advantage of an opportunity or avoid a risk.

### 1.3.13 Unstructured Data Analytics

An iterative process of scanning and tagging to add hooks to unstructured data, to allow filtering and linking to related structured data.

### 1.3.14 Operational Analytics

BI or streaming analytics.

### 1.3.15 Data Visualisation

The process of interpreting concepts, ideas and facts by using pictures of graphical representations.

### 1.3.16 Data Mashups

Combine data and internet-based services to create visualisation for insight and analysis

# 2 Activities

## 2.1 Define Big Data Strategy and Business Needs

Must be aligned with overall business strategy:

- What problems does the organisation need analytics to solve?
- What data sources to use or acquire?
- The timeliness and scope of the data to provision
- The impact on and relation to other data structures
- Influences to existing modelled data

## 2.2 Choose Data Sources

Big Data comes from many internal and external sources. Evaluate the Quality and reliability, and know its origin (provenance), format, what elements represent, how it connects to other data, and how frequently it will be updated.

Review available data sources, processes that create the data, and manage the plan for new sources.

- **Foundational data:** e.g. POS in sales analysis
- **Granularity:** Most granular form is ideal
- **Consistency:** Select data that appears appropriately and consistently across visualisations
- **Reliability:** Use trusted, authoritative sources
- **Inspect / profile new sources:** Test changes before adding new data sets

Risks:

- Privacy
- Filters which may introduce bias

## 2.3 Acquire and Ingest Data Sources

Capture critical Metadata about the source (origin, size, currency and additional knowledge about the content) when ingesting data sources into the Big Data environment. Ingestion engines may profile the data. Provides information on how to integrate with other data sets, such as Master Data or historical warehouse data, also how to train and validate models.

## 2.4 Develop Data Hypotheses and Methods

Data science entails building statistical models that find correlations and trends within and between data sets to find insights within the data. Models should be tested for a range of outcomes. Models depend on the quality of input data.

## 2.5 Integrate / Align Data for Analysis

Preparing data for analysis involves understanding what is in the data and links between data from various sources.

- Daily data would have to be aggregated to link to monthly data.
- Common key
- Similarity and record linking algorithms

- Clustering used to determine groupings

### 2.6 Explore Data using Models

- **Populate the Predictive Model**
- **Train the Model:** Repeated runs of the model against the data resulting in changes to the model
- **Evaluate the Model:**
  - Evaluated and validated against training sets
  - Ethical training to remove the biases of the creators
- **Create Data Visualisations:** Ensure the visualisation addresses the audience.

### 2.7 Deploy and Monitor

- Expose Insights and Findings
- Integrate with additional Data Sources

# 3 Tools

### 3.1 MPP (Massively Parallel Processing) Shared-nothing Technologies and Architecture

A system that automatically distributes data and parallelises query workload across all available hardware.

Data is partitioned across multiple processing nodes each processing data locally. Communication is controlled by a master host and occurs over a network. There is no disk sharing or memory contention. Linearly scalable. Distribution of workload to processor level. Speeds up analytical functions.

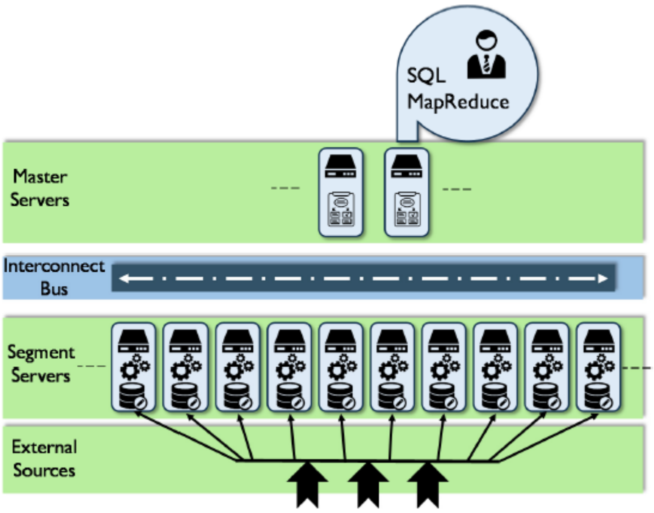


Figure 102 Columnar Appliance Architecture[95]

### 3.2 Distributed File-based Databases:

- Open source Hadoop stores files of any type
- Language used is called MapReduce:
  - **Map:** Identify and obtain data to be analysed
  - **Shuffle:** Combine data according to desired analytical patterns
  - **Reduce:** Remove duplication and perform aggregation to reduce the size of the data set to only what is required.

### 3.3 In-database Algorithms

Each processor in an MPP Shared-nothing platform can run queries independently.  Computation close to data reduces time for complex algorithms.

### 3.4 Big Data Cloud Solutions

Vendors provide cloud storage and enhancements

### 3.5 Statistical computing and graphical languages

R – scripting language with statistical computing and graphics.

### 3.6 Data visualisation tools

**Traditional visualisation tools:** both data and graphical

**Information graphics / Infographics:** Insights, changes in data over time.  Interactive, sophisticated analysis, adherence to visualisation best practices

## 4 Techniques

- **Analytic Modelling:** Different depths of analysis:
  - **Descriptive modelling:** Summarises data structures in a compact manner
  - **Explanatory modelling:** statistical models for testing a hypothesis
- **Big Data Modelling:** Data needs to be integrated, specified and managed by applying traditional Enterprise Architecture principles.

## 5 Implementation Guidelines

### 5.1 Strategy Alignment

Strategically aligned with business objectives.  Documents goals, approach and governance principles.  Strategy deliverables should account for managing:

- Information lifecycle
- Metadata
- Data Quality
- Data acquisition
- Data access and security
- Data governance
- Data privacy
- Learning and adoption
- Operations

### 5.2 Readiness Assessment / Risk Assessment

Critical success factors:

- **Business relevance:** Big Data/Data Science initiatives must enforce business function
- **Business readiness:**
  - Prepared for long term delivery?
  - Committed to establishing centres of excellence to sustain the product?
  - How broad is the skill gap?
- **Economic viability:** Assessment of ownership costs – buying vs leasing
- **Prototype:** Can the solution be prototyped?

### 5.3 Organisation and Cultural Change

Need a communications program to engage stakeholders and a centre of excellence to provide training, best practices, knowledge management and communication across developer, designer, analyst and data consumer communities.

Cross functional roles:

- **Big Data platform architect:** Hardware, OS, file systems and services
- **Ingestion architect:** Data analysis, systems of record, modelling
- **Metadata specialist:** Metadata interfaces, architecture and contents
- **Analytic Design Lead:** End user analytic design
- **Data Scientist:** Architectural and model design based on statistical knowledge

## 6 Big Data and Data Science Governance

The enterprise view of data should drive decisions on sourcing, sharing, Metadata, enrichment and access.

### 6.1 Visualisation Channels management

Alignment of the appropriate visualisation tools at the right level of complexity for the user community.

### 6.2 Data Science and Visualisation Standards

Vital for customer-facing and regulatory-facing content:

- Tools standards by analytic paradigm, user community, subject area
- Requests for new data
- Data set process standard
- Processes to avoid biased results:
    - Data inclusion and exclusion
    - Assumptions in the models
    - Statistical validity of results
    - Validity of interpretation of results
    - Appropriate methods applied

### 6.3 Data Security

- Agree upon levels of access for authorised personnel
- Mask data for those not authorised
- Use encryption for highly sensitive data
- Recombination measures the ability to reconstitute sensitive or private data and must be managed
- Outcomes of analysis may violate privacy

### 6.4 Metadata

Managed as part of data ingestion else data lake becomes a swamp.

### 6.5 Data Quality

Data Quality is a measure of deviation from expected result, the smaller the difference the higher the quality. Mature Big Data organisations scan data inputs using data quality tools to understand the information within:

- **Discovery:** Where information resided in the data set
- **Classification:** What type of information based upon standardised patterns
- **Profiling:** How data is populated and structured
- **Mapping:** What other data sets can be mapped to these values

## 6.6 Metrics

- **Technical Usage Metrics:**
    - Look for data hot spots to manage distribution and performance
    - Growth rates for capacity planning
- **Loading and scanning metrics:**
    - Ingestion rate
    - Interaction with the community
    - Provided by execution logs
- **Learnings and stories:**
    - Counts and accuracy of models and patterns
    - Revenue realised for identified opportunities
    - Cost reduction from avoiding identified threats