Architecture for Big Data and Data Engineering

Big Data & Analytics

Cisco | Networking Academy®
Mind Wide Open™

Sections & Objectives

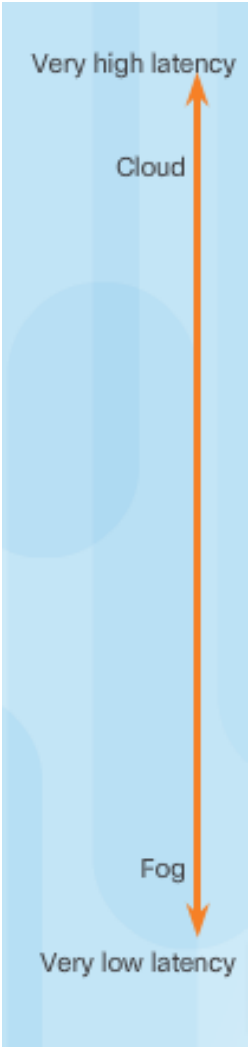
- Scaling Data Analytics
  - Explain how the virtualized data center supports Big Data and analytics.
- Introduction to Data Engineering
  - Explain the history, theory, concept, design, and barriers behind data engineering needs.
- The Big Data Pipeline
  - Explain how a big data pipeline supplies streaming IoT data for analysis.
- The Image Processing Labs
  - Analyze digital image data.

Scaling Data Analytics

Cisco | Networking Academy®
Mind Wide Open™

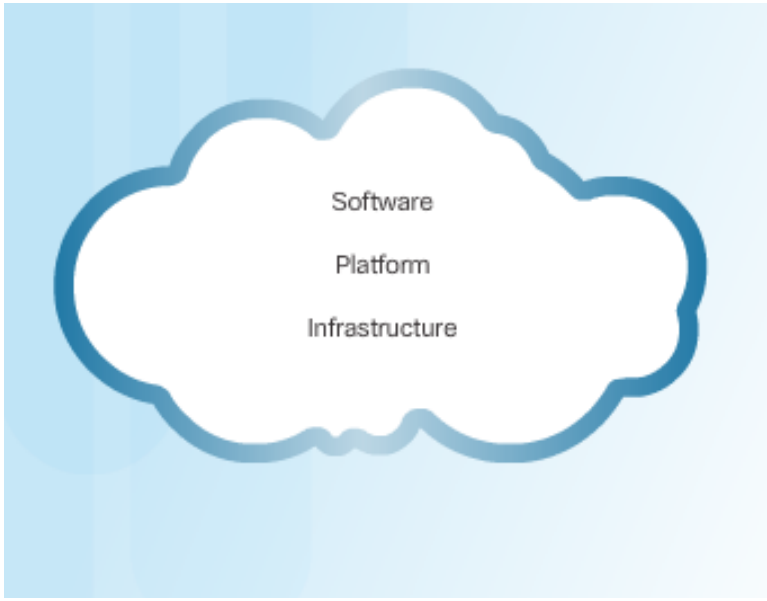
Edge Analytics and Cloud Analytics

- Transforming data into valuable insights requires computing and storage capacity.
- Device-Network-Cloud - all data points collected by sensors are sent directly to the cloud for storage and processing. This is what happens with most of the wearables used to track fitness activities.
- Device-Gateway-Network-Cloud - when the numbers of sensors increase, or when the processing of the data requires a much shorter response time, data can be processed very near the source of its creation on the gateway or other intermediate places on the network. Known as fog computing.



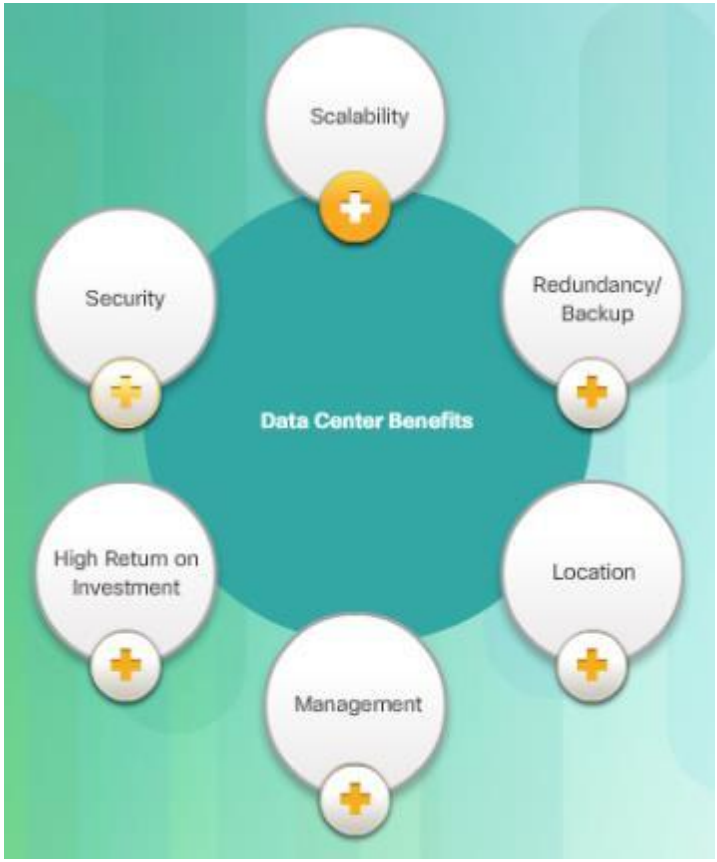
Data Centers and Cloud Computing

- Cloud Computing supports the four V's of Big Data: Volume, Variety, Velocity, Veracity
- Enterprise access to data anywhere anytime
- Pay-as-you-go model where you only subscribe to services that are needed
- Reduces costs by not having to purchase costly hardware or physical infrastructure
- Scalable computer storage and processing
- The 3 Main Cloud Services are:
  - SaaS – Software as a service
  - Paas – Platform as a service
  - IaaS – Infrastructure as a service



Benefits of a Data Center

- Some organizations create and maintain their own data centers in-house
- Other organizations rent data center servers at co-location facilities (colos)
- Other organizations use public, cloud-based services like Amazon Web Services, Microsoft Azure, Rackspace, and Google.
- Data centers provide:
  - Scalability,
  - Redundancy/Backup,
  - Location,
  - Management,
  - High return on investment,
  - Security

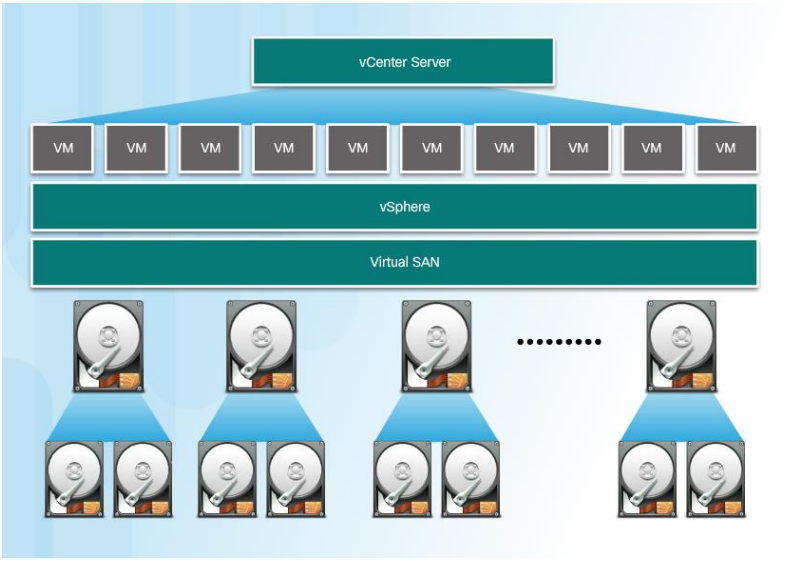


What is Virtualization?

- Virtualization separates the OS from the hardware.
- A hypervisor is software that creates and runs virtual machine (VM) instances.
- Containers are a specialized “virtual area”.

The Virtualized Data Center

- Data centers use virtualization to cut costs and expand offerings as cloud providers.
- Storage virtualization combines physical storage from multiple network storage devices into what appears to be a single storage device.
- Network virtualization (NV) is the creation of virtual networks within a virtualized infrastructure.



Introduction to Data Engineering

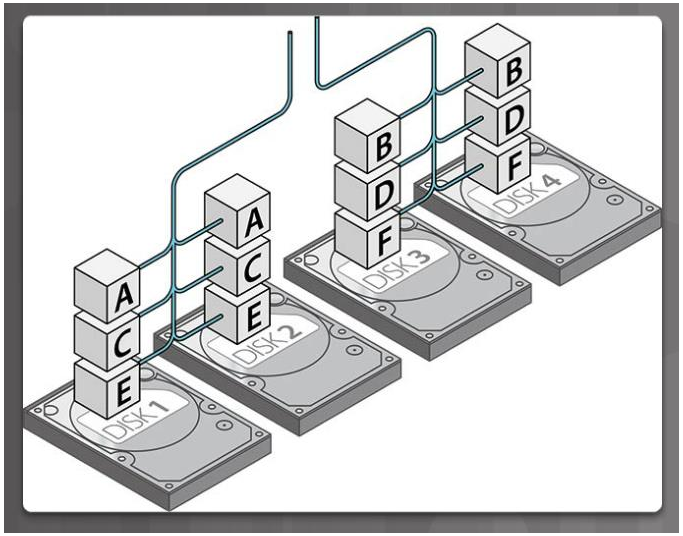
Cisco | Networking Academy®
Mind Wide Open™

What is Data Engineering?

- Data engineering typically involves a business-related, computer-based information system where information (data) is captured or generated, processed, stored, distributed, and analyzed.
- The ability to capture data and analyze it in a meaningful way is typically done with a database and database management system.
- The relational database emerged around the same time as the personal computer revolution.
  - The relational database and the structured query language (SQL) programming language are the foundation of the relational database management system (RDMS).
- The emergence of the Web 2.0, E-commerce and Google made it obvious that the relational database could not handle the volume and speed of web requests and searches.
- Non-relational databases like NoSQL and Object databases were created to meet the demands of the modern Web.
- Google helped pioneer the emergence of Big Data by openly publishing a paper on MapReduce and distributed processing and storage.

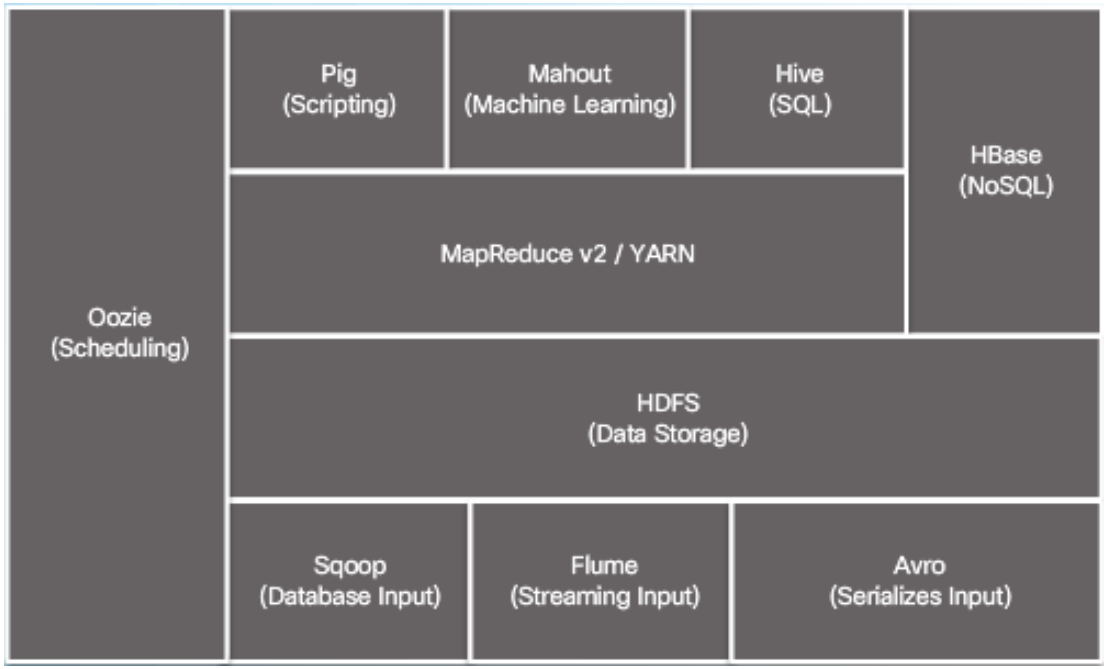
Big Data Systems

- Scalability is the ability to scale both data storage as well as data processing.
- Speed and availability are the primary concern for many companies working with Big Data.
- Fault tolerance is similar to availability in that a company's business needs to be constantly online and available 24/7.



What is Hadoop?

- The Hadoop Distributed File System (HDFS) is a redundant filesystem that stores data by distributing it across many computers.
- MapReduce is a distributed processing framework for parallelizing algorithms across large numbers of commodity servers.
- Hadoop is not a single application but an ecosystem of applications all working together.

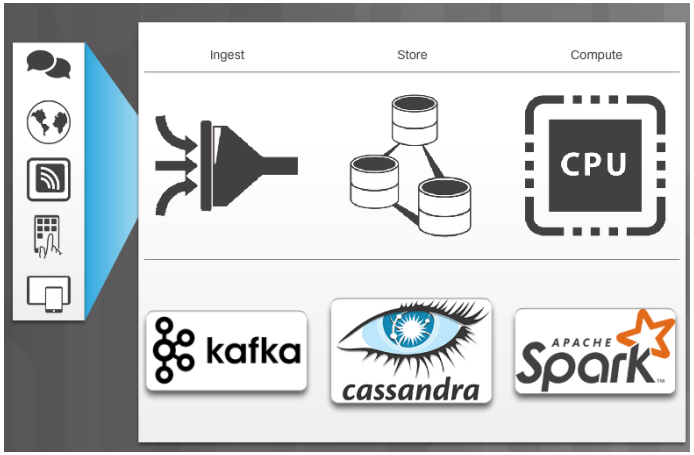


6.3 The Big Data Pipeline

Cisco | Networking Academy®
Mind Wide Open™

Data Ingestion

- The big data pipeline consists of: data ingestion, data storage, and data processing.
- To ingest data in real-time, a distributed streaming platform such as Kafka must be used.
- What makes Kafka different than traditional message brokers is the use of transaction logs.



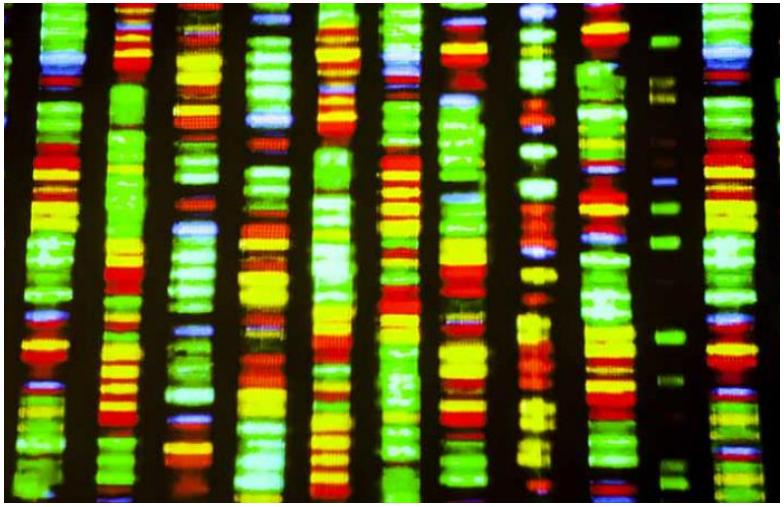
Data Storage

- Big Data generates vast amounts of data that must be stored.
- Cassandra is an open-source NoSQL distributed database management system.
- Cassandra uses the Cassandra File System (CFS).
- With the CFS, analytic metadata is stored in a keyspace.

The Big Data Pipeline

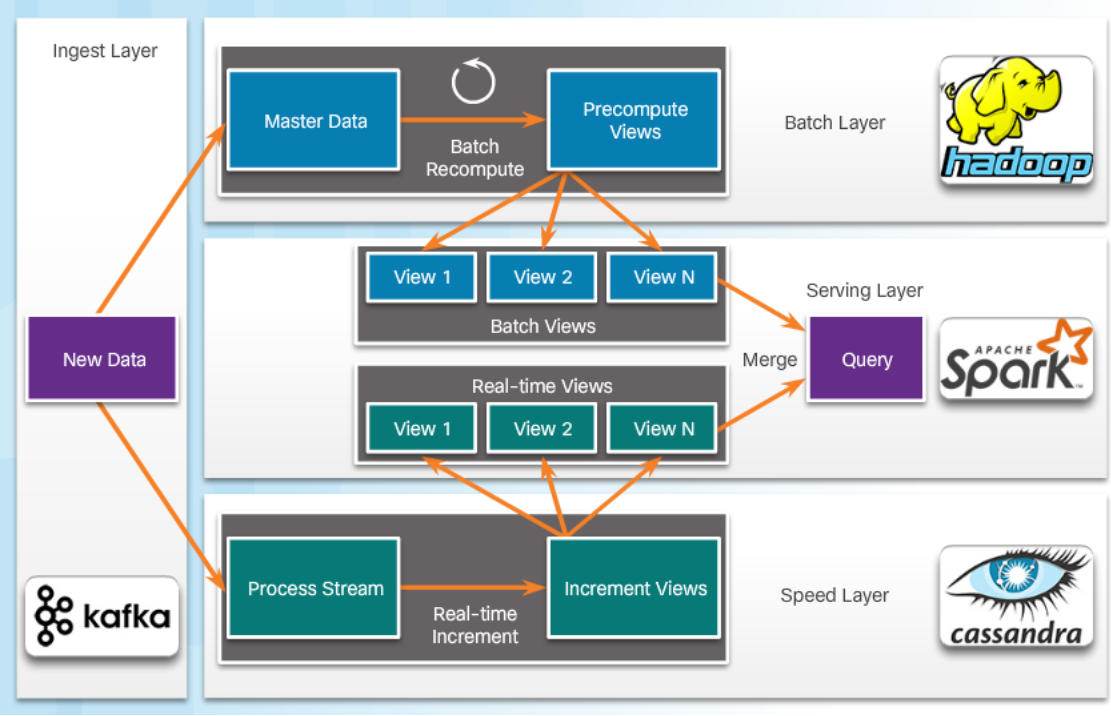
Compute

- The size of the data sets being used in many different fields is a challenge for Big Data.
- Spark is an open-source, distributed data processing engine used for Big Data.
- Spark is able to run right on top of an Hadoop instance, using HDFS for storage and YARN for cluster management.



The Lambda Architecture

- Lambda is a data processing architecture that uses both stream processing and batch processing to get accurate views of both “live” data and batch data.



6.4 The Image Processing Lab

Cisco | Networking Academy®
Mind Wide Open™

Digital Images as Data

- Data also includes media, such as images, video, and sound, as data.

Chapter Summary

Cisco | Networking Academy®
Mind Wide Open™

Summary

- Virtualized data center supports Big Data and analytics.
- With fog computing data can be processed almost immediately after it is generated.
- Data centers are centralized locations containing large amounts of computing and networking equipment.
- Virtualization separates the OS from the hardware.
- Storage virtualization combines physical storage from multiple network storage devices into what appears to be a single storage device.
- Network virtualization (NV) is the creation of virtual networks within a virtualized infrastructure.

Summary

- Data engineering involves a business-related, computer-based information system where information (data) is captured or generated, processed, stored, distributed, and analyzed.
- Scalability means designing a solution that can meet the exponential growth demands of large companies.
- The Hadoop Distributed File System (HDFS) is the filesystem where Hadoop stores data.
- MapReduce is a distributed processing framework for parallelizing algorithms across large numbers of commodity servers.
- Kafka is used to pipe real-time streaming data between different systems and applications.

Summary

- Cassandra uses the Cassandra File System (CFS) which is is not a master-slave architecture like HDFS.
- Cassandra is an open-source NoSQL distributed database management system.
- Spark is an open-source, distributed data processing engine used for Big Data jobs.
- Lambda is a data processing architecture that uses both stream processing and batch processing to get accurate views of both “live” data and batch data.
- In the digital age, media is numeric data also. It is represented by ones and zeros as digital data.