Chapter 6: A Big Data és Data Engineering architektúrája

Big Data & Analytics

Cisco | Networking Academy®
Mind Wide Open™

Chapter 6 - Sections & Objectives

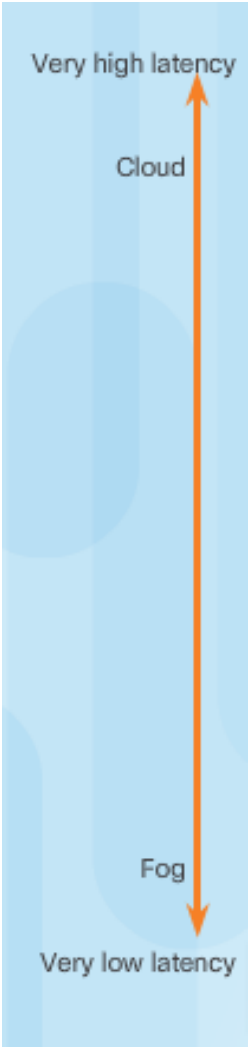
- 6.1 Az adatelemzés skálázása
  - Magyarázza el, hogyan támogatja a virtualizált adatközpont a Big Data-t és az analitikát.
- 6.2 Bevezetés az adatfeldolgozásba
  - Az adatechnikai igények mögött álló történelem, elmélet, koncepció, tervezés és akadályok magyarázata..
- 6.3 The Big Data Pipeline
  - Magyarázza el, hogy egy nagy adatpipeline hogyan szolgáltatja a streaming IoT-adatokat elemzésre..
- 6.4 A képfeldolgozó laborok
  - Digitális képi adatok elemzése.

6.1 Az adatelemzés skálázása

Cisco | Networking Academy®
Mind Wide Open™

Edge analitika és felhőelemzés

- Az adatok értékes meglátásokká alakítása számítási és tárolási kapacitást igényel..
- Eszköz-hálózat-felhő - az érzékelők által gyűjtött összes adatpontot közvetlenül a felhőbe küldik tárolásra és feldolgozásra. Ez történik a legtöbb fitnesztevékenységek nyomon követésére használt viselhető eszközzel..
- Device-Gateway-Network-Cloud - amikor az érzékelők száma növekszik, vagy amikor az adatok feldolgozása sokkal rövidebb válaszidőt igényel, az adatok feldolgozása nagyon közel a keletkezésük forrásához az átjárón vagy a hálózat más köztes helyein történhet. Ködszámításként ismert.

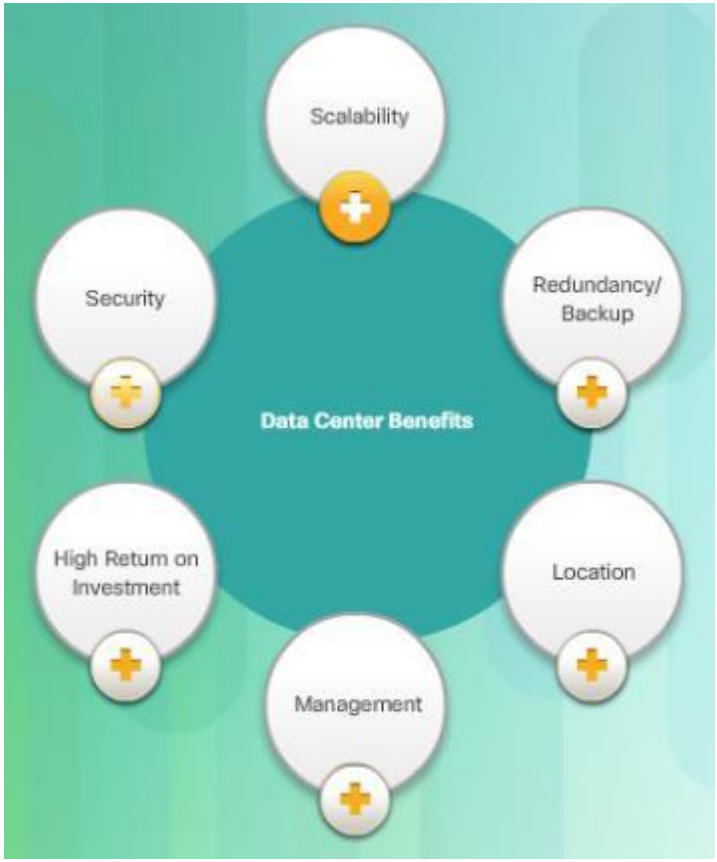


Data Centers and Cloud Computing

- A felhőalapú számítástechnika támogatja a Big Data négy V-jét: Volume, Variety, Velocity, Veracity
- Vállalati hozzáférés az adatokhoz bárhol és bármikor
- Pay-as-you-go modell, ahol csak a szükséges szolgáltatásokra fizet elő.
- Csökkenti a költségeket, mivel nem kell költséges hardvert vagy fizikai infrastruktúrát vásárolni.
- Skálázható számítógépes tárolás és feldolgozás.
- A 3 fő felhőszolgáltatás a következő:
  - SaaS – Software as a service
  - Paas – Platform as a service
  - IaaS – Infrastructure as a service

Benefits of a Data Center

- Vannak cégek, amelyek saját adatközpontokat hoznak létre és tartanak fenn házon belül..
- Más cegek adatközponti szervereket bérelnek ko-lokációs létesítményekben (colos)..
- Más cégek olyan nyilvános, felhőalapú szolgáltatásokat használnak, mint az Amazon Web Services, a Microsoft Azure, a Rackspace és a Google.
- Data centers provide:
  - Scalability,
  - Redundancy/Backup,
  - Location,
  - Management,
  - High return on investment,
  - Security

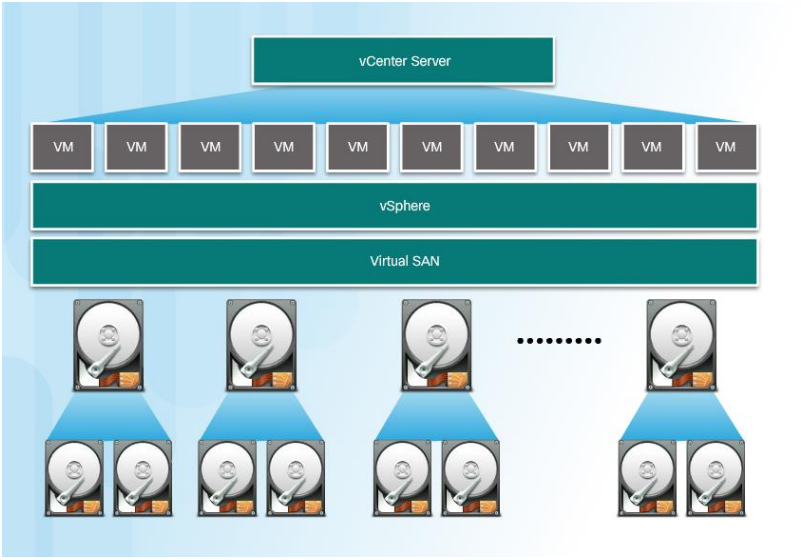


What is Virtualization?

- A virtualizáció elválasztja az operációs rendszert a hardvertől.
- A hypervisor egy olyan szoftver, amely virtuális gép (VM) példányokat hoz létre és futtat..
- A konténerek egy speciális "virtuális terület”.

The Virtualized Data Center

- Az adatközpontok virtualizációval csökkentik a költségeket és felhőszolgáltatóként bővítik kínálatukat.
- Storage virtualization combines physical storage from multiple network storage devices into what appears to be a single storage device.
- A hálózati virtualizáció (NV) virtuális hálózatok létrehozása egy virtualizált infrastruktúrán belül..



6.2 Introduction to Data Engineering

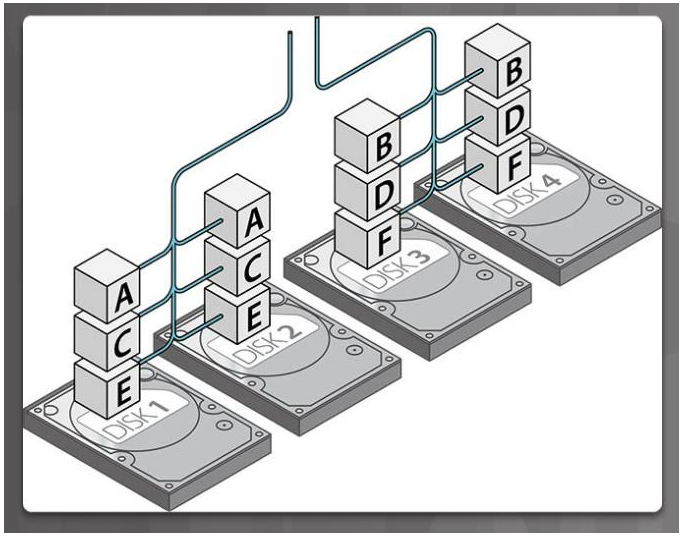
Cisco | Networking Academy®
Mind Wide Open™

What is Data Engineering?

- Az adatmérnökség jellemzően olyan üzleti vonatkozású, számítógép-alapú információs rendszert foglal magában, ahol az információkat (adatokat) rögzítik vagy generálják, feldolgozzák, tárolják, elosztják és elemzik.
- Az adatok rögzítésének és értelmes elemzésének képessége jellemzően adatbázis és adatbázis-kezelő rendszer segítségével valósul meg..
- A relációs adatbázis a személyi számítógépek forradalmával egy időben jelent meg..
  - A relációs adatbázis és a strukturált lekérdezési nyelv (SQL) programozási nyelv a relációs adatbázis-kezelő rendszer relational database management system (RDMS) alapja.
- A Web 2.0, az e-kereskedelem és a Google megjelenése nyilvánvalóvá tette, hogy a relációs adatbázisok nem képesek kezelni a webes kérések és keresések mennyiségét és sebességét..
- A nem-relációs adatbázisok, mint a NoSQL és az objektum adatbázisok a modern web igényeinek kielégítésére jöttek létre..

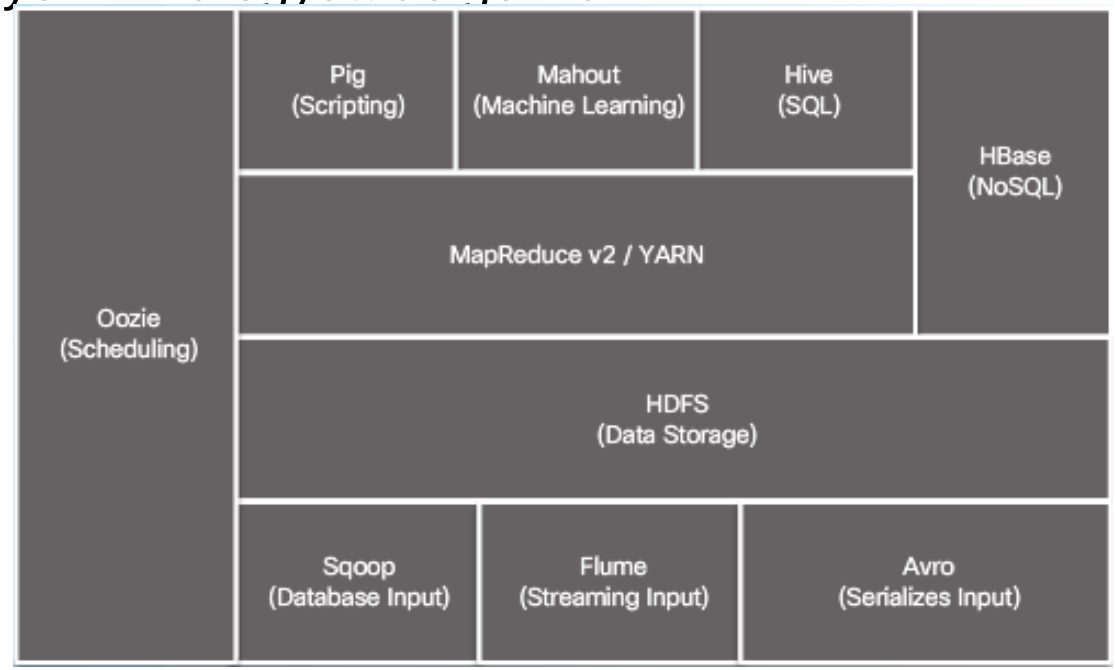
Big Data Systems

- A skálázhatóság az adattárolás és az adatfeldolgozás skálázhatósága.
- A sebesség és a elérhetőség az elsődleges szempont sok Big Data-val dolgozó vállalat számára...
- A hibatűrés annyiban hasonlít a elérhetőséghez, hogy a vállalat üzletének folyamatosan online és a nap 24 órájában elérhetőnek kell lennie..



What is Hadoop?

- A Hadoop Distributed File System (HDFS) egy redundáns fájlrendszer, amely az adatokat több számítógépen elosztva tárolja..
- A MapReduce egy elosztott feldolgozási keretrendszer algoritmusok párhuzamosítására nagyszámú alapvető szerveren.
- A Hadoop nem egyetlen alkalmazás, hanem alkalmazások ökoszisztémája, amelyek mind együtt dolgoznak..

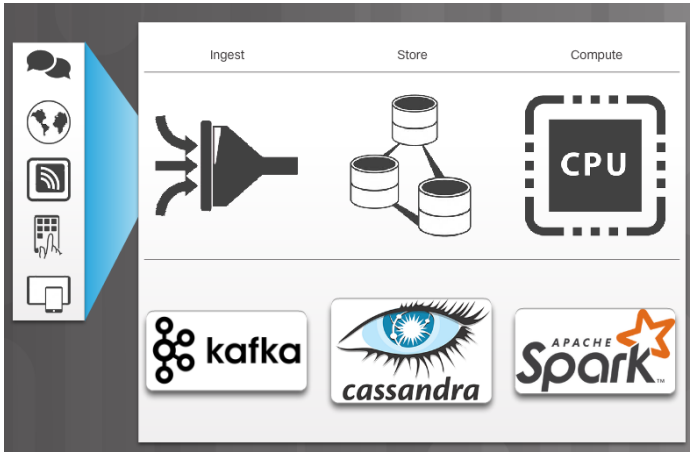


6.3 The Big Data Pipeline

Cisco | Networking Academy®
Mind Wide Open™

Data Ingestion

- A big data pipeline a következőkből áll: adatgyűjtés, adattárolás és adatfeldolgozás..
- A valós idejű adatbevitelhez olyan elosztott streaming platformot kell használni, mint a Kafka..
- A Kafka a tranzakciós naplók használata miatt különbözik a hagyományos üzenetközvetítőktől..



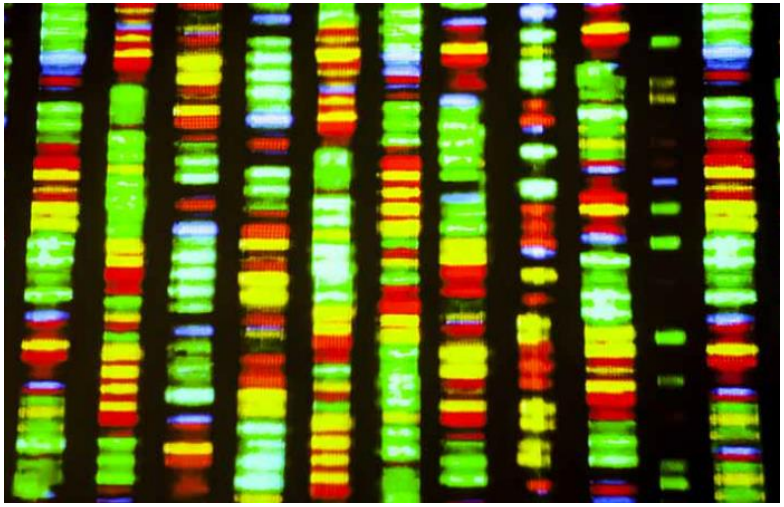
Data Storage

- A Big Data hatalmas mennyiségű adatot generál, amelyet tárolni kell..
- A Cassandra egy nyílt forráskódú NoSQL elosztott adatbázis-kezelő rendszer.
- A Cassandra a Cassandra fájlrendszert (CFS) használja.
- A CFS esetében az analitikus metaadatokat egy kulcstőtérben tárolják..

The Big Data Pipeline

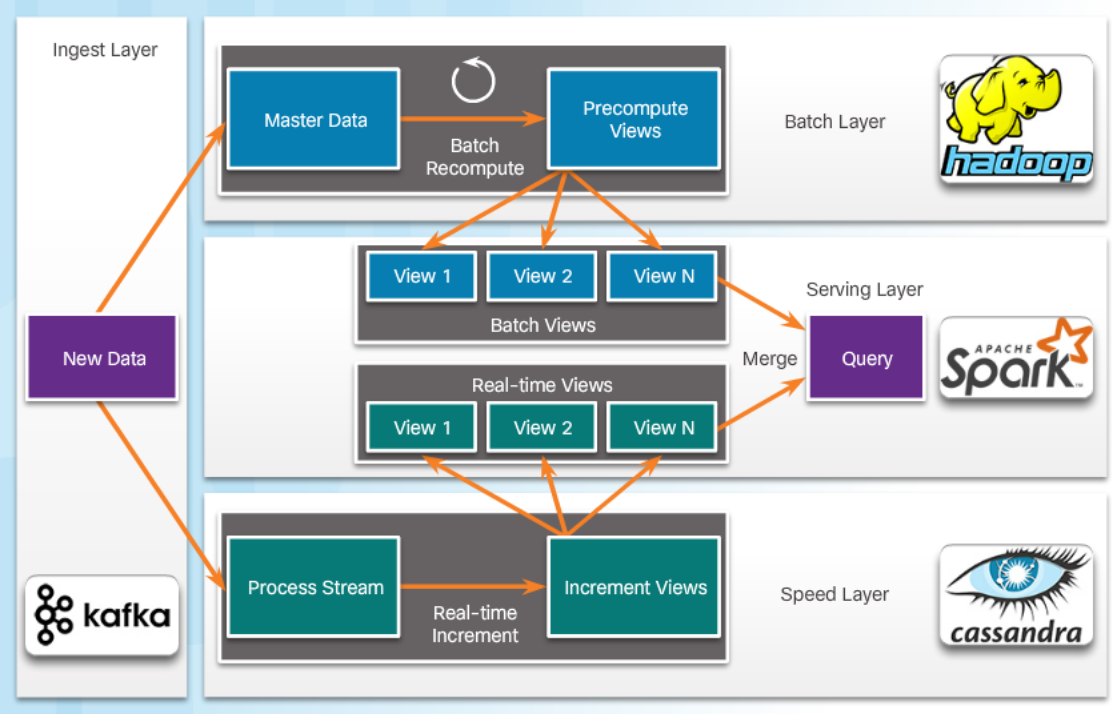
Compute

- A számos különböző területen használt adathalmazok mérete kihívást jelent a Big Data számára.
- A Spark egy nyílt forráskódú, elosztott adatfeldolgozó motor, amelyet a Big Data.
- A Spark képes közvetlenül egy Hadoop-példányon futni, HDFS-t használva a tároláshoz és YARN-t a klaszterek kezeléséhez..



The Lambda Architecture

- A Lambda egy adatfeldolgozási architektúra, amely folyamfeldolgozást és kötegelt feldolgozást egyaránt használ, hogy pontos képet kapjon mind az "élő" adatokról, mind a kötegelt adatokról..



6.4 The Image Processing Lab

Cisco | Networking Academy®
Mind Wide Open™

The Image Processing Lab

Digital Images as Data

- Az adatok közé tartoznak az olyan adathordozók is, mint a képek, videók és hangok..

6.5 Chapter Summary

Cisco | Networking Academy®
Mind Wide Open™

Summary

- A virtualizált adatközpont támogatja a Big Data-t és az analitikát.
- A ködszámítással az adatok szinte azonnal feldolgozhatók a létrehozásuk után..
- Az adatközpontok olyan központosított helyek, amelyek nagy mennyiségű számítástechnikai és hálózati berendezést tartalmaznak..
- A virtualizáció elválasztja az operációs rendszert a hardvertől.
- A hálózati virtualizáció (NV) virtuális hálózatok létrehozása egy virtualizált infrastruktúrán belül..

Summary

- Az adatmérnökség olyan üzleti vonatkozású, számítógép-alapú információs rendszert foglal magában, ahol az információkat (adatokat) rögzítik vagy generálják, feldolgozzák, tárolják, elosztják és elemzik..
- A skálázhatóság olyan megoldás megtervezését jelenti, amely képes megfelelni a nagyvállalatok exponenciális növekedési igényeinek..
- A Hadoop Distributed File System (HDFS) az a fájlrendszer, amelyben a Hadoop az adatokat tárolja.
- A Kafka a különböző rendszerek és alkalmazások közötti valós idejű streaming adatok továbbítására szolgál..

Summary

- A Cassandra a Cassandra File System (CFS) rendszert használja, amely nem master-slave architektúra, mint a HDFS..
- A Cassandra egy nyílt forráskódú NoSQL elosztott adatbázis-kezelő rendszer..
- A Spark egy nyílt forráskódú, elosztott adatfeldolgozó motor, amelyet Big Data feladatokhoz használnak.
- A Lambda egy adatfeldolgozási architektúra, amely folyamfeldolgozást és kötegelt feldolgozást egyaránt használ, hogy pontos képet kapjon mind az "élő" adatokról, mind a kötegelt adatokról..
- A digitális korban a média is numerikus adat. Digitális adatként egyesek és nullák jelölik..