Introducció al BigData

Ll. Gesa

Arquitectura i tecnologies software

UAB, 2017

Index

Alternatives a Hadoop & Spark

Alternatives a Hadoop & Spark

Hadoop & Spark les més conegudes, per la seva estesa implantació i antiguitat, però no les úniques.

- Storm (Apache). Event Real-Time oriented.
- DataTorrent RTS (Apache). Real-Time oriented.
- Cluster Mapreduce (Chitika)
- High Performance Computing Cluster (HPCC).
- Hydra (AddThis)
- Google DataFlow (Next slides)
- Amazon Web Services , AWS
- Dryad (Microsoft)

Alternatives a Hadoop & Spark

Hadoop & Spark les més conegudes, per la seva estesa implantació i antiguitat, però no les úniques.

- Storm (Apache). Event Real-Time oriented.
- DataTorrent RTS (Apache). Real-Time oriented.
- Cluster Mapreduce (Chitika)
- High Performance Computing Cluster (HPCC).
- Hydra (AddThis)
- Google DataFlow (Next slides)
- Amazon Web Services , AWS
- Dryad (Microsoft)

http://www.fromdev.com/2015/03/hadoop-alternatives.html

BigML

Que és?

BigML, Inc. És una companyia que ofereix solucions de Machine learning per manipular i analitzar qualsevol tipus de dada (sense límit de tamany de dades)

BigML

Que és?

BigML, Inc. És una companyia que ofereix solucions de Machine learning per manipular i analitzar qualsevol tipus de dada (sense límit de tamany de dades)

història

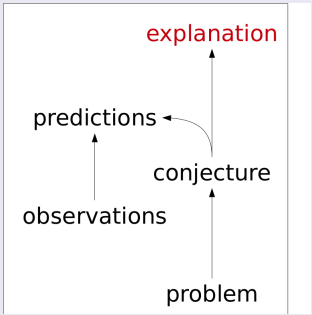
- 2011 - Fundació
- 2012 - Crea estrategia de mercat enfocat en els Models Predictius.
- 2013 - Rep $1.3M de finanzament
- 2013 - BigML es converteix en una eina de Cloud Prediction
- 2014 - Release de Advanced Predictive Modeling platform
- 2016 - Telefònica Open Future i BigML s'uneixen per crear PreSeries: Una eina de 'Early Stage Investment'

BigML

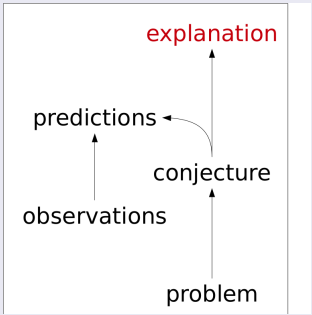
Productes

- **BigML.io** - Cloud REST API per integrar Machine Learning i models predictius en projectes fent servir la infrastructura BigML.
- **BigMLer** - Client de Command line tool d'acces a l'interface REST.
- **The BigML PredictServer** - Imatge de servidor amb BigML tecnologia per desplegar-se en els serveis clouds de cada projecte.
- **Flatline** - A Lisp-like llenguatge d'acces a la infrastructura BigML.
- **WhizzML** - Llenguatge de programació d'alt nivell per el domini especific de Machine Learning.

Raonament Inductiu amb Machine Learning



Raonament Inductiu amb Machine Learning



Implementació

Arbres de decisió

BigML

Tecnologia

- Basat en application/json.
- RESTFul Services (HTTP verbs GET, POST, PUT, DELETE plus status notifications: queued, in-progress, finished(error))
- Nginx server
- MongoDB for metadata
- FileSystem propi sobre infrastructura aliena (per exemple S3 Amazon).
- Infrastructura propia (ni Haddop, Spark..)

Arquitectura

Patró MVC: Front-End, Middle-End, Back-End

Index

Enginyeria de software en Bigdata: Introducció

Grans Volums de dades necessita nous paradigmes de proces:

- Hardware
- Sistema d'arxius
- Base de dades
- Nous Algoritmes

Hardware

Super-Computers ?

Systema d'Arxius

Systema d'Arxius

HDFS

Systema d'Arxius

HDFS

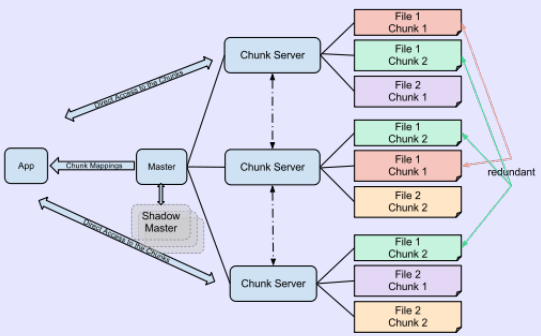
GoogleFS

Actualment millorat, amb nom Colossus

Systema d'Arxius

HDFS

GFS



GoogleFS

Actualment

Systema d'Arxius

HDFS

GoogleFS

Actualment millorat, amb nom Colossus

S3 (Amazon)

99.99% durability and 99.99% availability

Systema d'Arxius

HDFS

GoogleFS

Actualment millorat, amb nom Colossus

S3 (Amazo

99.99% durability and 99.99% availability

Amazon Data Architecture: Paper

Dynamo: Amazon's Highly Available Key-value Store

Systema d'Arxius

HDFS

GoogleFS

Actualment millorat, amb nom Colossus

S3 (Amazon)

99.99% durability and 99.99% availability

Google Cloud Storage

Sobre Colossus

Systema d'Arxius

HDFS

GoogleFS

Actualment

Cloud Platforms

Amazon i Google ofereixen el seus serveis de storage per sí sols o amb l'entorn/plataforma per sobre: Amazon ASW (EC2), o Google Cloud.
Creació de maquines virtuals on-demand.

S3 (Amazon)

99.99% durability and 99.99% availability

Google Cloud Storage

Sobre Colossus

Systema d'Arxius

HDFS

GoogleFS

Actualment millora

Cloud Computing

S3 (Amazon)

99.99% durability and 99.99% availability

Google Cloud Storage

Sobre Colossus

Systema d'Arxius

HDFS

GoogleFS

Actualment millora

S3 (Amazon)

99.99% durability a

Google Cloud Storage

Sobre Colossus

Cloud Computing

Infrastructure as a Service (IaaS)
Platform as a Service (PaaS)
Software as a Service (SaaS)

Systema d'Arxius

HDFS

GoogleFS

Actualment millora

S3 (Amazon)

99.99% durability a

Google Cloud Storage

Sobre Colossus

...

Molts més.

Cloud Computing

Infrastructure as a Service (IaaS)
Platform as a Service (PaaS)
Software as a Service (SaaS)

Base de Dades

Sql ? No-SQL ?

Tema 2

Sql ? No-SQL ?

Tema 2

MySQL issue:9544

[31 Mar 2005 22:10] Thad Welch

Description:

I'm trying to create indexes on a table with 308 million rows. It took ~20 minutes to load the table but 10 days to build indexes on it. The table's MYD file is 3.2G and its MYI file is 7.7G.

BigData: Design Patterns?

Grans volums de dades necessiten noves formes d'afrontar els problemes

Cassos d'ús

Exemples

BigData: Design Patterns?

Grans volums de dades necessiten noves formes d'afrontar els problemes

Cassos d'ús

Exemples

- Prediccions de consum elèctric.

BigData: Design Patterns?

Grans volum

Cassos d'ús

Exemples

- Predicci

Consum elèctric

Les empreses de serveis públics han llançat comptadors intel·ligents per mesurar el consum d'aigua, gas i electricitat a intervals regulars d'una hora o menys. Aquests mesuradors intel·ligents generen grans volums de dades d'interval que necessita ser analitzat. Aquestes empreses executen grans, cars i complicats sistemes per generar energia. Cada element d'aquestes infraestructures inclou sofisticats sensors que monitoritzen la tensió, corrent, freqüència i altres característiques. Per guanyar eficiència operativa, l'empresa ha de supervisar les dades lliurades pel sensor. Una solució en BigData pot analitzar la generació d'energia (alimentació) i les dades de consum d'energia (demanda) i trobar les millors mesures per maximitzar-ne tot el procés.

BigData: Design Patterns?

Grans volums de dades necessiten noves formes d'afrontar els problemes

Cassos d'ús

Exemples

- Prediccions de consum elèctric.
- Sentiment Social vers una marca.

BigData: Design Patterns?

Grans volum

Cassos d'ús

Exemples

- Predicci
- Sentime

Sentiment Social

Els departaments de màrqueting utilitzen els post de Twitter per dur a terme anàlisis dels sentiments per determinar quins usuaris estan dient què sobre l'empresa i els seus productes o serveis, especialment després del llançament d'un nou producte. El 'sentiment' del client ha d'estar integrat amb les dades del perfil del client per obtenir resultats significatius. Amb la retroalimentació de dades dels clients pot fer variar els resultats d'acord amb el temps i potser la pròpia demografia dels clients.

BigData: Design Patterns?

Grans volums de dades necessiten noves formes d'afrontar els problemes

Cassos d'ús

Exemples

- Prediccions de consum elèctric.
- Sentiment Social vers una marca.
- Log Monitoring.

BigData: Design Patterns?

Grans volum

Cassos d'ús

Exemples

- Predicci
- Sentimen
- Log Monitoring.

Log Monitoring

Els departaments de TI estan recorren a solucions de BigData per analitzar registres de l'aplicació per obtenir una perspectiva que pot millorar el rendiment del sistema. Un problema típic en aquest àmbit és que els arxius de registre de diverses ampliacions poden tenir formats diferents i molt cop poc clars: S'han de normalitzar abans de poder-los utilitzar.

BigData: Design Patterns?

Grans volums de dades necessiten noves formes d'afrontar els problemes

Cassos d'ús

Exemples

- Prediccions de consum elèctric.
- Sentiment Social vers una marca.
- Log Monitoring.
- Detecció de frau.

Grans volum

Cassos d'ús

Exemples

- Predicci
- Sentime
- Log Mo
- Detecci

Detecció del Frau

La gestió de frau prediu la probabilitat que una determinada transacció o compte de client està experimentant el frau. Solucions per analitzar les transaccions en temps real i generar recomanacions per a l'acció immediata, que és fonamental per aturar el frau de tercers, el frau de primera part, i el mal ús deliberat de privilegis dels comptes. Les solucions es dissenyen típicament per detectar i prevenir el frau i el risc de tipus innombrables a través de múltiples indústries, incloent:

- Crèdit i targetes de pagament de dèbit el frau
- frau en els seus comptes de dipòsit
- frau tècnic
- Deute incobrable
- frau d'atenció mèdica
- Assegurances de responsabilitat civil de frau
- frau de la remuneració del treballador
- el frau d'assegurances
- frau de telecomunicacions

BigData: Design Patterns?

Grans volums de dades necessiten noves formes d'afrontar els problemes

Cassos d'ús

Exemples

- Prediccions de consum elèctric.
- Sentiment Social vers una marca.
- Log Monitoring.
- Detecció de frau.
- Reconeixement Facial per interacció Home-Maquina.

BigData: Design Patterns?

Grans volum

Reconeixement facial

Les empreses poden utilitzar la tecnologia de reconeixement facial en combinació amb una foto de les xarxes socials per fer ofertes personalitzades als clients basats en el comportament de compra i la ubicació. Aquesta capacitat podria tenir un gran impacte en les empreses. Aquesta faceta té ramificacions greus de la privacitat de les dades.

Cassos d'ús

Exemples

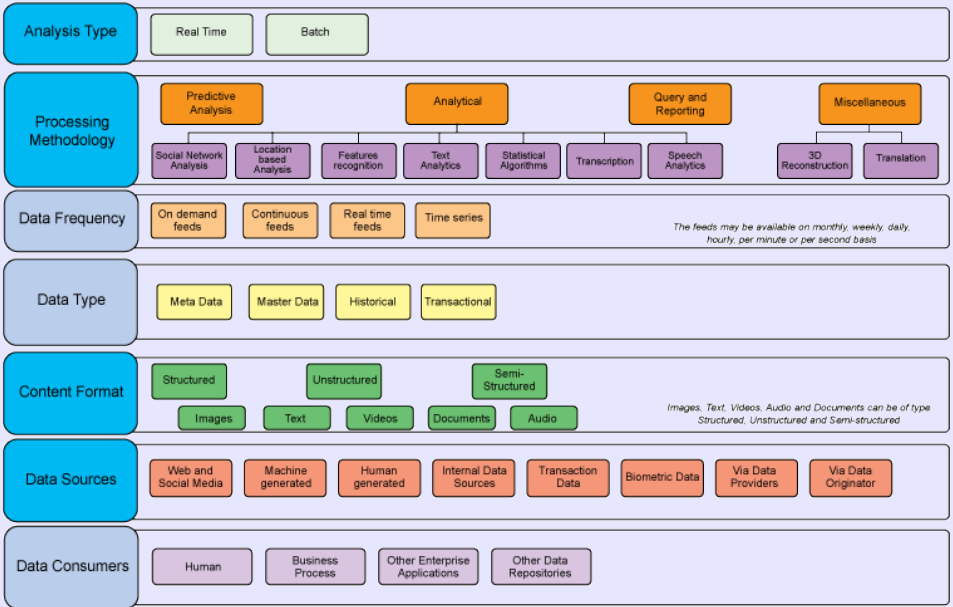
- Predicci
- Sentiment social vers una marca.
- Log Monitoring.
- Detecció de frau.
- Reconeixement Facial per interacció Home-Maquina.

Parametrització del problema

Un cop amb el problema sobre la taula cal analitzar l'origen de les dades i certs aspectes del proces buscat:

- El format del contingut.
- El tipus (transaccional, històric, real-time(sensors)..)
- La frequència amb que hi ha dades noves.
- La frequència amb que calen resultats, com s'ha de processar les dades: Temps Real, casi Temps-Real, en offline.

Classificació



Un cop a
proces bu
- El fc
- El ti
- La fr
- La fr
  Tem

lel

si

Parametrització del problema

Concepte: ETL

Extract, Transform and Load (extreure, Transformar i carregar (ETL)) és el procés que permet a les organitzacions moure dades des de múltiples fonts, reformateixarlos i netejar-los, i carregar-los en una altra base de dades per analitzar , o en un altre sistema operacional per donar suport a un procés de negoci.

Un cop amb el pr :s aspectes del proces buscat:

- El format del
- El tipus (trans
- La frequència
- La frequència amb que caien resultats, com s'ha de processar les dades: Temps Real, casi Temps-Real, en offline.

Patrons

- Patrons estructurals

Patrons

- Patrons estructurals
  - Real-Time Streaming
  - Near Real-Time streaming
  - Lambda Architecture
  - Kappa Architecture
  - Data-Lake

Patrons: Problemes típics, solucions típiques

Patrons

- Patrons estructurals
  - Real-Time Streaming
  - Near Real-Time streaming
  - Lambda Architecture
  - Kappa Architecture
  - Data-Lake
- Patrons Funcionals

Patrons: Problemes típics, solucions típiques

Patrons

- Patrons estructurals
  - Real-Time Streaming
  - Near Real-Time streaming
  - Lambda Architecture
  - Kappa Architecture
  - Data-Lake
- Patrons Funcionals
  - Stream Joins
  - Top N (trending)
  - Rolling Windows
  - Data Historic

Patrons: Problemes típics, solucions típiques

Patrons

- Patrons estructurals
  - Real-Time Streaming
  - Near Real-Time streaming
  - Lambda Architecture
  - Kappa Architecture
  - Data-Lake
- Patrons Funcionals
  - Stream Joins
  - Top N (trending)
  - Rolling Windows
  - Data Historic
- Gestió de dades

Patrons: Problemes típics, solucions típiques

Patrons

- Patrons estructurals
  - Real-Time Streaming
  - Near Real-Time streaming
  - Lambda Architecture
  - Kappa Architecture
  - Data-Lake
- Patrons Funcionals
  - Stream Joins
  - Top N (trending)
  - Rolling Windows
  - Data Historic
- Gestió de dades
  - External lookup
  - Responsible Shuffing
  - Out-of-Sequence

Patrons: Problemes típics, solucions típiques

Patrons

- Patrons estructurals
  - Real-Time Streaming
  - Near Real-Time streaming
  - Lambda Architecture
  - Kappa Architecture
  - Data-Lake
- Patrons Funcionals
  - Stream Joins
  - Top N (trending)
  - Rolling Windows
  - Data Historic
- Gestió de dades
  - External lookup
  - Responsible Shuffing
  - Out-of-Sequence
- Patrons de seguretat

Patrons: Problemes típics, solucions típiques

Patrons

- Patrons estructurals
  - Real-Time Streaming
  - Near Real-Time streaming
  - Lambda Architecture
  - Kappa Architecture
  - Data-Lake
- Patrons Funcionals
  - Stream Joins
  - Top N (trending)
  - Rolling Windows
  - Data Historic
- Gestió de dades
  - External lookup
  - Responsible Shuffing
  - Out-of-Sequence
- Patrons de seguretat
  - Message Encryptation
  - Authorized Access
  - Secure Cluster Authentication

Patrons: Problemes típics, solucions típiques

Patrons

- Patrons estructurals
  - Real-Time Streaming
  - Near Real-Time streaming
  - Lambda Architecture
  - Kappa Architecture
  - Data-Lake
- Patrons Funcionals
  - Stream Joins
  - Top N (trending)
  - Rolling Windows
  - Data Historic
- Gestió de dades
  - External lookup
  - Responsible Shuffing
  - Out-of-Sequence
- Patrons de seguretat
  - Message Encryptation
  - Authorized Access
  - Secure Cluster Authentication
- Patrons de consum

Patrons: Problemes típics, solucions típiques

Patrons

- Patrons estructurals
  - Real-Time Streaming
  - Near Real-Time streaming
  - Lambda Architecture
  - Kappa Architecture
  - Data-Lake
- Patrons Funcionals
  - Stream Joins
  - Top N (trending)
  - Rolling Windows
  - Data Historic
- Gestió de dades
  - External lookup
  - Responsible Shuffing
  - Out-of-Sequence
- Patrons de seguretat
  - Message Encryptation
  - Authorized Access
  - Secure Cluster Authentication
- Patrons de consum
  - Visualització
  - Descobriment Ad-Hoc
  - Notificacions

Patrons: Problemes típics, solucions típiques

Patrons

- Patrons estructurals
  - Real-Time Streaming
  - Near Real-Time streaming
  - Lambda Architecture
  - Kappa Architecture
  - Data-Lake
- Patrons Funcionals
  - Stream Joins
  - Top N (trending)
  - Rolling Windows
  - Data Historic
- Gestió de dades
  - External lookup
  - Responsible Shuffing
  - Out-of-Sequence
- Patrons de seguretat
  - Message Encryptation
  - Authorized Access
  - Secure Cluster Authentication
- Patrons de consum
  - Visualització
  - Descobriment Ad-Hoc
  - Notificacions
- Patrons d'access

Patrons: Problemes típics, solucions típiques

Patrons

- Patrons estructurals
  - Real-Time Streaming
  - Near Real-Time streaming
  - Lambda Architecture
  - Kappa Architecture
  - Data-Lake
- Patrons Funcionals
  - Stream Joins
  - Top N (trending)
  - Rolling Windows
  - Data Historic
- Gestió de dades
  - External lookup
  - Responsible Shuffing
  - Out-of-Sequence
- Patrons de seguretat
  - Message Encryptation
  - Authorized Access
  - Secure Cluster Authentication
- Patrons de consum
  - Visualització
  - Descobriment Ad-Hoc
  - Notificacions
- Patrons d'access
  - Web i social Media
  - Device

Patrons: Problemes típics, solucions típiques

Patrons

- Patrons estructurals
  - Real-Time Streaming
  - Near Real-Time streaming
  - Lambda Architecture
  - Kappa Architecture
  - Data-Lake
- Patrons Funcionals
  - Stream Joins
  - Top N (trending)
  - Rolling Windows
  - Data Historic
- Gestió de dades
  - External lookup
  - Responsible Shuffing
  - Out-of-Sequence
- Patrons de seguretat
  - Message Encryptation
  - Authorized Access
  - Secure Cluster Authentication
- Patrons de consum
  - Visualització
  - Descobriment Ad-Hoc
  - Notificacions
- Patrons d'access
  - Web i social Media
  - Device
- Patrons de resolució

Patrons: Problemes típics, solucions típiques

Patrons

- Patrons estructurals
  - Real-Time Streaming
  - Near Real-Time streaming
  - Lambda Architecture
  - Kappa Architecture
  - Data-Lake
- Patrons Funcionals
  - Stream Joins
  - Top N (trending)
  - Rolling Windows
  - Data Historic
- Gestió de dades
  - External lookup
  - Responsible Shuffing
  - Out-of-Sequence
- Patrons de seguretat
  - Message Encryptation
  - Authorized Access
  - Secure Cluster Authentication
- Patrons de consum
  - Visualització
  - Descobriment Ad-Hoc
  - Notificacions
- Patrons d'access
  - Web i social Media
  - Device
- Patrons de resolució
  - Map-Reduce

Patrons: Problemes típics, solucions típiques

Patrons

- Patrons estructurals
  - Real-Time Streaming
  - Near Real-Time streaming
  - Lambda Architecture
  - Kappa Architecture
  - Data-Lake
- Patrons Funcionals
  - Stream Joins
  - Top N (trending)
  - Rolling Windows
  - Data Historic
- Gestió de dades
  - External lookup
  - Responsible Shuffing
  - Out-of-Sequence
- Patrons de seguretat
  - Message Encryptation
  - Authorized Access
  - Secure Cluster Authentication
- Patrons de consum
  - Visualització
  - Descobriment Ad-Hoc
  - Notificacions
- Patrons d'access
  - Web i social Media
  - Device
- Patrons de resolució
  - Map-Reduce
  - Stream Processing/Pipeline/DataFlow

Patrons: Problemes típics, solucions típiques

Estructural: Lamnda. Cas d'ús

La resposta d'un sistema està directament lligada a les dades mes recents obtingudes (sensors, peticions de clients), no obstant l'històric té un gran valor per acabar de parametritzar la resposta. El volum de dades és gran (Gb diaris).

Patrons: Problemes típics, solucions típiques

Estructural: Lamnda. Cas d'ús

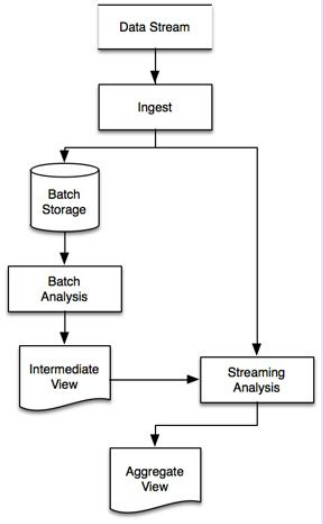
La resposta d'un sistema està directament lligada a les dades mes recents obtingudes (sensors, peticions de clients), no obstant l'històric té un gran valor per acabar de parametritzar la resposta. El volum de dades és gran (Gb diaris).

Aplicar Lamnda

- Les dades s'injecten paral.lelament al sistema d'Stream analítics i al arxiu de dades.
- El sistema offline funciona contínuament actualitzant unes vistes intermitges.
- El sistema streaming en temps real combina les ultimes dades generades per el sistema offline amb les dades fresques més actuals

Patrons: Problemes

Lambda



Estructural: Lamnda.

La resposta d'un sistema
peticions de clients), no
resposta. El volum de da

dades mes recents obtingudes (sensors,
lor per acabar de parametritzar la

Aplicar Lamnda

- Les dades s'injecten
- El sistema offline fu
- El sistema streaming
  offline amb les dade

eam analítics i al arxiu de dades.
nt unes vistes intermitges.
imes dades generades per el sistema

Patrons: Problemes típics, solucions típiques

Estructural: Data Lake. Cas d'ús

Es vol analitzar l'impacte en xarxes socials, blog i webs d'un cert producte. Per tan, les dades originals seran de formats diversos. Tampoc és té molt clar el tipus d'anàlisis que s'haurà de realitzar i si totes les dades capturades seran útils.

Patrons: Problemes típics, solucions típiques

Estructural: Data Lake. Cas d'ús

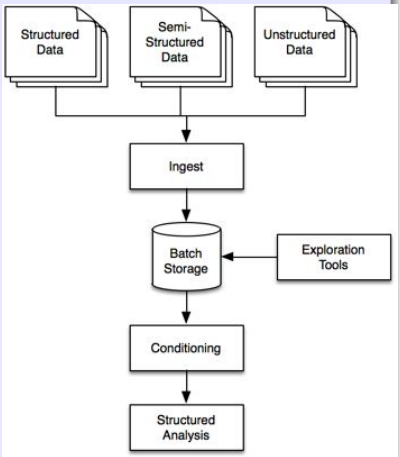
Es vol analitzar l'impacte en xarxes socials, blog i webs d'un cert producte. Per tan, les dades originals seran de formats diversos. Tampoc és té molt clar el tipus d'anàlisis que s'haurà de realitzar i si totes les dades capturades seran útils.

Aplicar Data Lake

- Totes les dades s'injecten en una base de dades inicial que suporti contingut sense estructura (per exemple HDFS directament).
- Es produeix un anàlisi inicial per identificar les dades útils i el procediment a seguir.
- S'acondicionen les dades (s'estructuren) i es realitza un anàlisis final
- Opcionalment les dades estructurades es poden re-injectar a la base de dades o a una altra.

Patrons: Problemes

Lake



Estructural: Data Lake

Es vol analitzar l'impacte un cert producte. Per tan, les dades originals seran de format lar el tipus d'anàlisis que s'haurà de realitzar i si totes les dad

Aplicar Data Lake

- Totes les dades s'inj l que suporti contingut sense estructura (per exen
- Es produeix un anàl útils i el procediment a seguir.
- S'acondicionen les d un anàlisis final
- Opcionalment les dades estructurades es poden re injectar a la base de dades o a una altra.

Patrons: Problemes típics, solucions típiques

D'accés: Device-generated

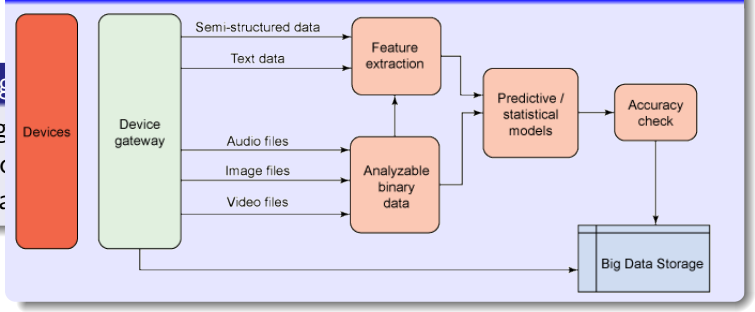
Les dades son generades per dispositius sensors, de diversa indole: amb informació meteorològica, elèctrica, dades sobre la contaminació. Les dades poden ser fotos, vídeos, text i/o informació binaria.

Patrons: Problemes típics, solucions típiques

D'accés: Device-g

Les dades son g
meteorològica, elèc
i/o informació bina

Device-generated data pattern



nb informació
s, vídeos, text

Patrons: Problemes típics, solucions típiques

Suport Teòric

http://www.ibm.com/developerworks/library/bd-archpatterns4/index.html

Patrons: Problemes típics, solucions típiques

De Resolució: Map-Reduce

Ja vist. Resolució en 2 fases: Map com fase inicial, Reduce com fase final.

Patrons: Problemes típics, solucions típiques

De Resolució: Stream Processing

Tantes fases com faci falta, múltiples operadors/transformacions. **RDD** de **Spark** o **Google DataFlow**.

Patrons: Problemes típics, solucions típiques

De Resolució: Stream Processing

Tantes fases com faci falta, múltiples operadors/transformacions. **RDD** de **Spark** o **Google DataFlow**.

Google Dataflow

La documentació de Google DataFlow ofereix un guiatge de que fer servir, per via de 4 preguntes: Què?, on?, Quan?, Com?:

- Quins resultats es calculen? Es respon amb quina transformació cal.
- On es calculen els resultats? Gestió del cluster.
- Quan es calculen? RealTime, watermarks, triggers.
- Com es calculen? Trandormacions i agrupaments.

Google DataFlow

Què és?

- Data processing system: batch and streaming
- Set of SDKs
- Google Cloud Platform managed services:
  - Google Compute Engine (VMs)
  - Google Cloud Storage (r/w data)
  - BigQuery (r/w data)

Google DataFlow

Model de programació

- Basat en patró Pipeline
- Estructura **PCollection** : Conjunt de dades dins el pipeline
- Transformacions **PTransforms**: Qualsevol process sobre les dades
- Pipeline I/O - Serialització de dades

Google DataFlow

Model de programació

- Basat en patró Pipeline
- Estructura **PCollection** : Conjunt de dades dins el pipeline
- Transformacions **PTransforms**: Qualsevol process sobre les dades
- Pipeline I/O - Serialització de dades

PCollection

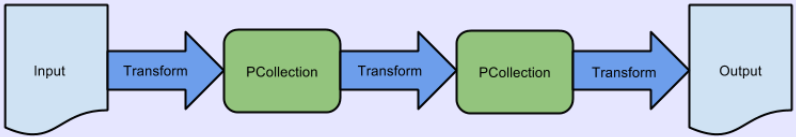
- Represent una data en el pipeline
- Potencialment inlimitat (stream)
- Serialitzable, immutable, no access aleatori als seus elements.
- Deferred data (potser no s'ha computat)
- Fortament lligat al Windowing i/o triggers

Google DataFlow

Model de programació

- Basat e
- Estructu
- Transfo
- Pipeline

Pipeline Lineal



PCollection

- Represent una data en el pipeline
- Potencialment inlimitat (stream)
- Serialitzable, immutable, no access aleatori als seus elements.
- Deferred data (potser no s'ha computat)
- Fortament lligat al Windowing i/o triggers

Google DataFlow

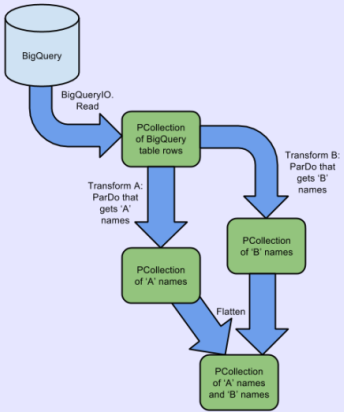
Model de programació

- Basat en patr
- Estructura **PC**
- Transformacic
- Pipeline I/O -

PCollection

- Represent una
- Potencialment
- Serialitzable, i
- Deferred data
- Fortament llig

Pipeline de branca

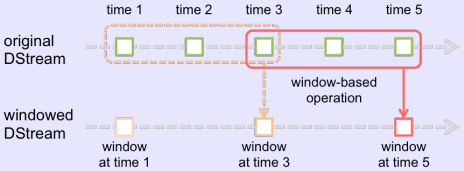


Google DataFlow

Model de programació

- Basat en patró Pipe
- Estructura **PCollect**
- Transformacions **PT**
- Pipeline I/O - Serial

Windowing



PCollection

- Represent una data en el pipeline
- Potencialment inlimitat (stream)
- Serialitzable, immutable, no access aleatori als seus elements.
- Deferred data (potser no s'ha computat)
- Fortament lligat al Windowing i/o triggers

PTransforms

- Computacions matemàtiques
- Conversió de formats
- Agrupacions
- Filtratge
- Reduccions
- Sobrecàrrega

MP vs Spark vs Google Dataflow : Gestió de Recursos

Google DataFlow és un entorn d'execució completament sota demanda. Específicament, quan s'executa un treball els recursos s'assignen sobre la demanda per només aquest treball. No hi ha intercanvi / contenció de recursos a través de llocs de treball. En comparació amb **Spark** o **MapReduce** el més habitual és implementar un clúster de nodes X i després enviar els treballs i després ajustar els recursos de node a través de llocs de treball. El model **Google DataFlow** està orientat a una gestió menys manual.

MP vs Spark vs Google Dataflow : Gestió de Recursos

Google DataFlow és un entorn d'execució completament sota demanda. Específicament, quan s'executa un treball els recursos s'assignen sobre la demanda per només aquest treball. No hi ha intercanvi / contenció de recursos a través de llocs de treball. En comparació amb **Spark** o **MapReduce** el més habitual és implementar un clúster de nodes X i després enviar els treballs i després ajustar els recursos de node a través de llocs de treball. El model **Google DataFlow** està orientat a una gestió menys manual.

MP vs Spark vs Google Dataflow : Interactivitat

Google DataFlow actualment no és dinàmic: Un cop llançat els process, el sistema evoluciona sol i mostra els resultats finals. **Spark** proporciona mecanismes per consultar dinàmicament el process menstres aquest s'executa. **Map-Reduce** també sols prorciona resultats al final.

MP vs Spark vs Google Dataflow : Model de programació

Google DataFlow i **Spark** tenen una visí més de llenguatge funcional, mentre **Map-Reduce** és més imperatiu classic. Mentres **Spark** suporta varis llenguatges de base, **Google DataFlow** sols **Java**^{tm}i recentment **Python**. **Map-Reduce** te suport de varis llenguatges per via de eines de 'traducció'.

Google DataFlow

MP vs Spark vs Google Dataflow : Model de programació

Google DataFlow i **Spark** tenen una visí més de llenguatge funcional, mentre **Map-Reduce** és més imperatiu classic. Mentres **Spark** suporta varis llenguatges de base, **Google DataFlow** sols **Java**^{tm}i recentment **Python**. **Map-Reduce** te suport de varis llenguatges per via de eines de 'traducció'.

MP vs Spark vs Google Dataflow : Streaming i Windowing

Google DataFlow es l'entorn que suporta millor les 2 filosofies de processament: Streaming (Temp-Real) o Windowing (Casi Temps real). **Spark** També les suporta en una eficiencia similar. **Map-Reduce** més encarat a process offline

Index