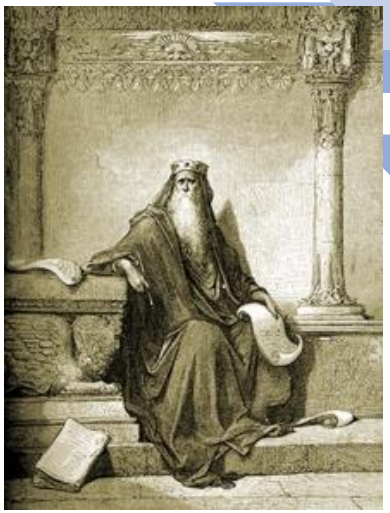
Conceptos Fundamentales de de Big Data

Computación de Alto Desempeño
Diego Alberto Rincón Yáñez MCSc
diego-rincon@javeriana.edu.co

Palabras de Koheleth Hijo de David, Rey en Jerusalem 200 D.C.

“ No hay nada Nuevo bajo el Sol!”

Ecclesiastes Chapter 1 verso 9



Diego Alberto Rincón Yáñez MCSc.
Twitter: @d1egoprog.

Antes de empezar……

¿Qué es Big Data?

Diego Alberto Rincón Yáñez MCSc.
Twitter: @d1egoprog.

¿De donde la moda del Big Data?

"Big Data is like teenage sex, everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it."

Dan Arieli,
Duke University Director
Center for Advanced Hindsight

Diego Alberto Rincón Yáñez MCSc.
Twitter: @d1egoprog.

¿Qué sabemos realmente de Big Data?

- Big Data es solo "data"
  - ¿Cual es la mejor forma de manejar nuestros datos?
- Big Data es una pieza en un rompecabezas.
  - ¿Cómo combinarla con la analítica existente para lograr un gran impacto?
- Big Data necesita unirse a la operación del negocio.
  - ¿Cómo podemos usar big data para crear mejores productos y servicios?

NO al botón de reinicio, no estamos iniciando desde cero.

Diego Alberto Rincón Yáñez MCSc.
Twitter: @d1egoprog.

¿Qué sabemos realmente de Big Data?

Nivel Organización

- Convertir la cultura de la organización en “data-driven”.
- Usar preguntas de negocio, no de tecnología para guiar la arquitectura.
- Incrementar la autogestión de datos, para fomentar la decisión a partir de datos (Empoderar al Usuario).
- Habilitar un ecosistema que permita a los equipos “nadar” en los datos para descubrir.
- Hacer los datos y los resultados tan accesibles a la gente como sea posible.

Diego Alberto Rincón Yáñez MCSc.
Twitter: @d1egoprog.

¿Qué sabemos realmente de Big Data?

Nivel Tecnología

- Construir un modelo estructurado del negocio.
- Descubrir información acerca de clientes y productos.
  - ERP + CRM +Sistemas Legados
  - Logs web.
  - Correos + Redes Sociales
  - Voz + Texto
  - BI (Business Intelligence)
- Asegurar la información como un activo, controlando el acceso a los datos.
  - Proteger privacidad
  - Cumplir con regulaciones

Diego Alberto Rincón Yáñez MCSc.
Twitter: @d1egoprog.

Agenda

- Conceptos Básicos
- Fases del Proceso
- Arquitecturas de Big Data
- Tecnologías de Big Data
- Definición de Proyecto
- Casos de Estudio

Diego Alberto Rincón Yáñez MCSc.
Twitter: @d1egoprog.

Big Data & Data Science

Diego Alberto Rincón Yáñez MCSc.
Twitter: @d1egoprog.

Balanceo de Términos

- Analítica (Des, Diag, Pred y Pres)
- Data Science vs Big Data
- Arquitectura (Sw, IT, BD, Empr)
- B.D. (Architech, Engineer, Scientist)
- CDO (Chief Data Officer)
- Business Intelligence
- KPI (Key Performance Indicators)
- Open (Data vs Access)
- Metadata
- ETL vs ELT
- OLTP vs OLAP
- Data (Warehouse, Marts)
- Paralelo vs Distribuido
- Stream vs batch
- Clúster (HA vs HPC vs HTC)
- SQL y NoSQL
- Hadoop (HDFS+YARN+MR)

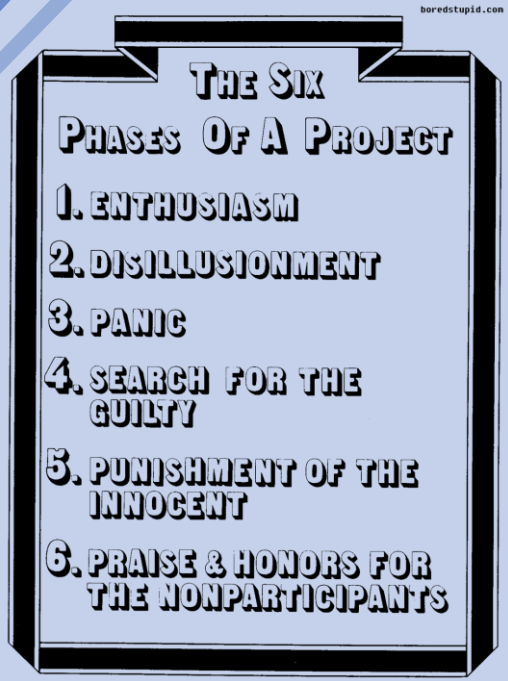
Diego Alberto Rincón Yáñez MCSc.
Twitter: @d1egoprog.

Big Data no es solo Hadoop

- Ejemplos donde Hadoop no es completamente aplicable
  - Cyber Seguridad, mercados de bolsa, información de sensores y tendencias en redes sociales (monitoreo en tiempo real).
  - Que pasaría si, su empresa tuviera muchos centros de datos, donde el ancho de banda es un problema.
  - Control de los datos/Seguridad (en sectores específicos).

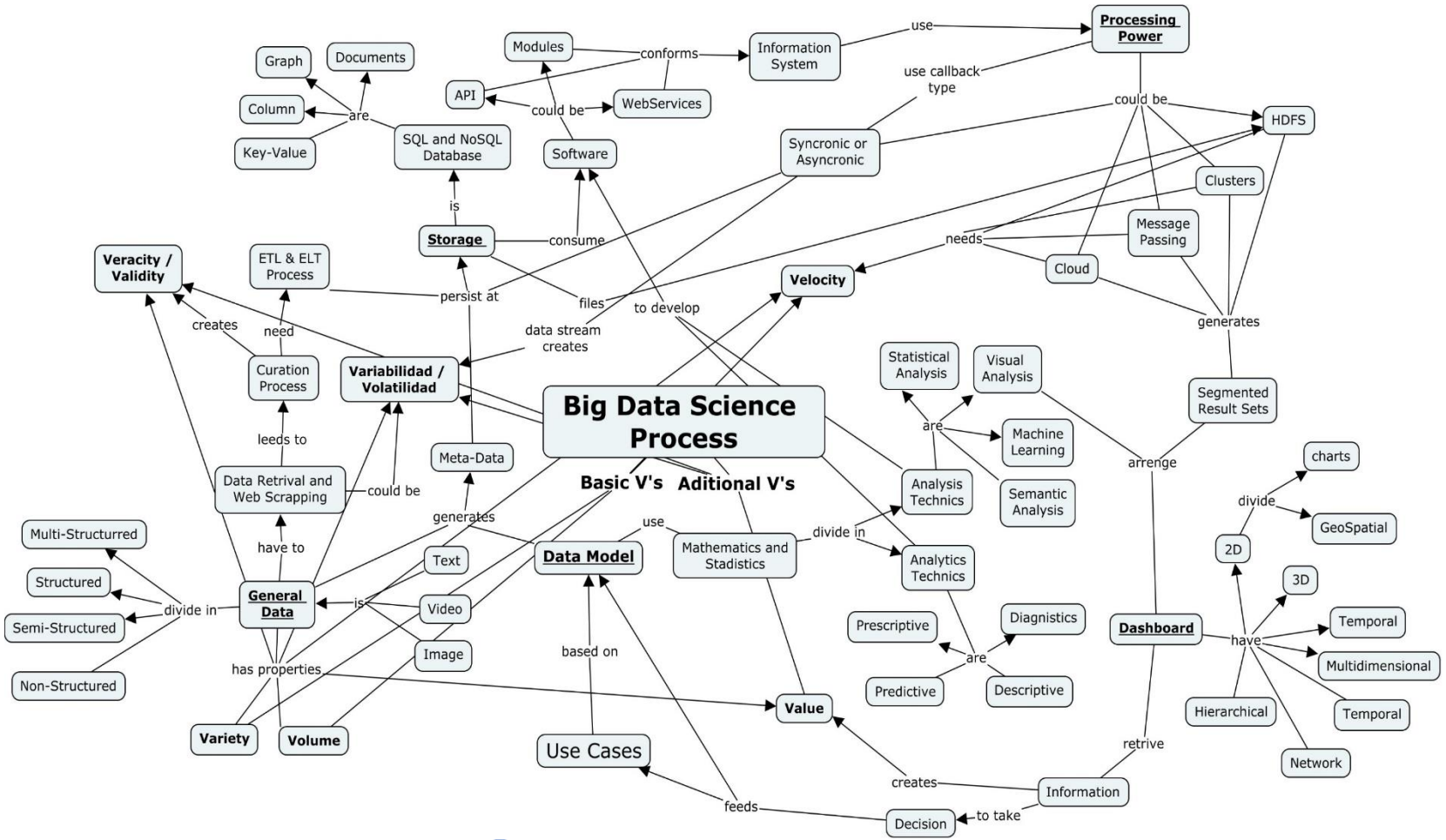
- Más detalles en el módulo más adelante.

Diego Alberto Rincón Yáñez MCSc.
Twitter: @d1egoprog.

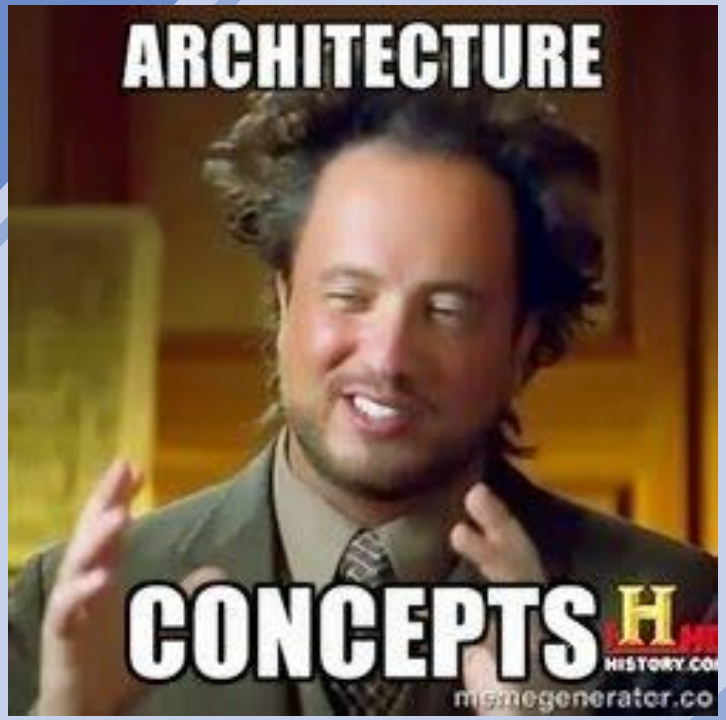


Fases del proceso

Diego Alberto Rincón Yáñez MCSc.
Twitter: @d1egoprog.



Diego Alberto Rincón Yáñez MCSc.
Twitter: @d1egoprog.



Conceptos de Arquitecturas de Big Data

Diego Alberto Rincón Yáñez MCSc.
Twitter: @d1egoprog.

Hoy Big Data es Procesamiento en Batch

Favorite Product List Changes

1.2.13	Add	iPAD 64GB
10.3.13	Add	Sony RX-100
11..3.13	Add	Canon GX-10
11.3.13	Remove	Sony RX-100
12.3.13	Add	Nikon S-100
14.4.13	Add	BoseQC-15
15.4.13	Add	MacBook Pro 15
20.4.13	Remove	Canon GX10

Raw information => data

derive

Current Favorite Product List

iPAD 64GB
Nikon S-100
BoseQC-15
MacBook Pro 15

derive

Current Product Count

4

Information => derived

Diego Alberto Rincón Yáñez MCSc.
Twitter: @d1egoprog.

Operacional

Características

- Producida día a día
- Alto Volumen
- Baja Latencia
- Operaciones CRUD pequeñas (OLTP)
- Ejemplos
  - Datos de Cliente
  - Inventario
  - Compras

Diego Alberto Rincón Yáñez MCSc.
Twitter: @d1egoprog.

Características

- Múltiples Dominios (Negocios)
- Decisiones de Negocio
- Ajustado a la imaginación del analista.
- Optimizado para Mining, Ad hoc, batch, etc (OLTP)
- Ejemplos
  - Segmentación de Clientes
  - Pronósticos
  - Cambios de hábitos

Analítica

Diego Alberto Rincón Yáñez MCSc.
Twitter: @d1egoprog.

Arquitecturas en Números

	Operacional	Analítica
Latencia	1ms a 100ms	1min a 100min
Concurrencia	1000 a 100000	1 a 10
Tipos de Acceso	Lect / Escr	Lectura
Consultas	Selectivas	No Selectivas
Alcance	Operaciones	Histórico
Tecnología	NoSQL	MPP y MapReduce
Usuario Final	Cliente	Científico de datos

Diego Alberto Rincón Yáñez MCSc.
Twitter: @d1egoprog.

Principios de Arquitecturas Orientadas a Datos

1. **Abrir la mente a los datos**, todos, no solo los grandes, es critico extraer valor de todos.
2. **Toda la tecnología es relevante,** Data warehouse, descubrimiento y HDFS, son complementarias y buscar integración con el resto de la arquitectura empresarial, procesamiento de eventos, visualización y analítica.
3. **Reemplazar la pila por una "cadena de suministro",** un sistema que de respuesta a los datos no puede ser una pila, sino ajustarse como una cadena.

Diego Alberto Rincón Yáñez MCSc.
Twitter: @d1egoprog.

Principios de Arquitecturas Orientadas a Datos

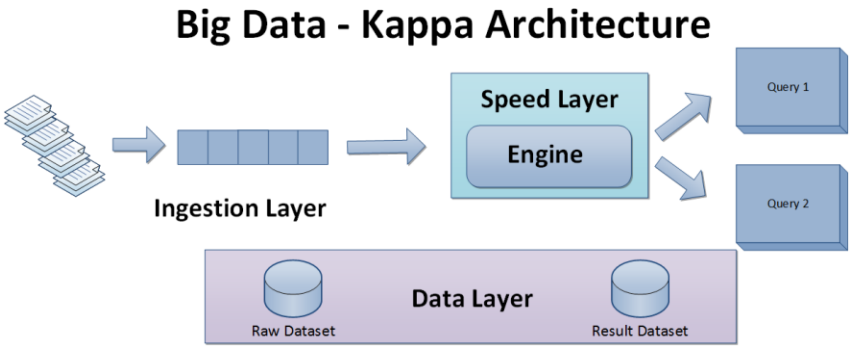
4. **Los consumidores de datos no son creados igual,** los usuarios tiene diferentes habilidades, permisos y formas de recibir la información. Entregar información de la forma que la consumen.
5. **Conserve lo que funciona,** Establecer políticas alrededor de gobierno, calidad, seguridad y accesibilidad de los datos.
6. **Empiece pequeño,** Diseñar proyectos pequeños que den resultados rápido, no gastar tiempo en áreas de poco valor.

Diego Alberto Rincón Yáñez MCSc.
Twitter: @d1egoprog.

Arquitecturas estándares para Big Data

Diego Alberto Rincón Yáñez MCSc.
Twitter: @d1egoprog.

Arquitectura Kappa

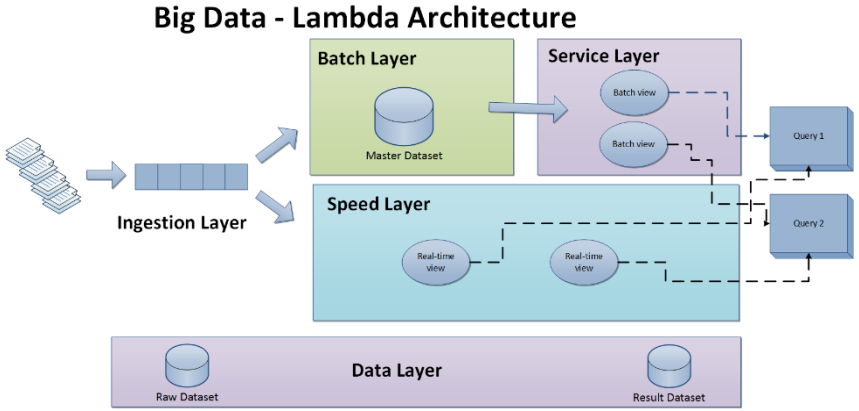


- la capa de datos
- la capa de velocidad

Diego Alberto Rincón Yáñez MCSc.
Twitter: @d1egoprog.

Arquitectura Lambda

- la capa de datos
- la capa batch
- la capa de servicio
- la capa de velocidad



Diego Alberto Rincón Yáñez MCSc.
Twitter: @d1egoprog.

Taxonomías de Arquitecturas de Big Data

¿Quiénes usan arquitecturas de referencia de big data?

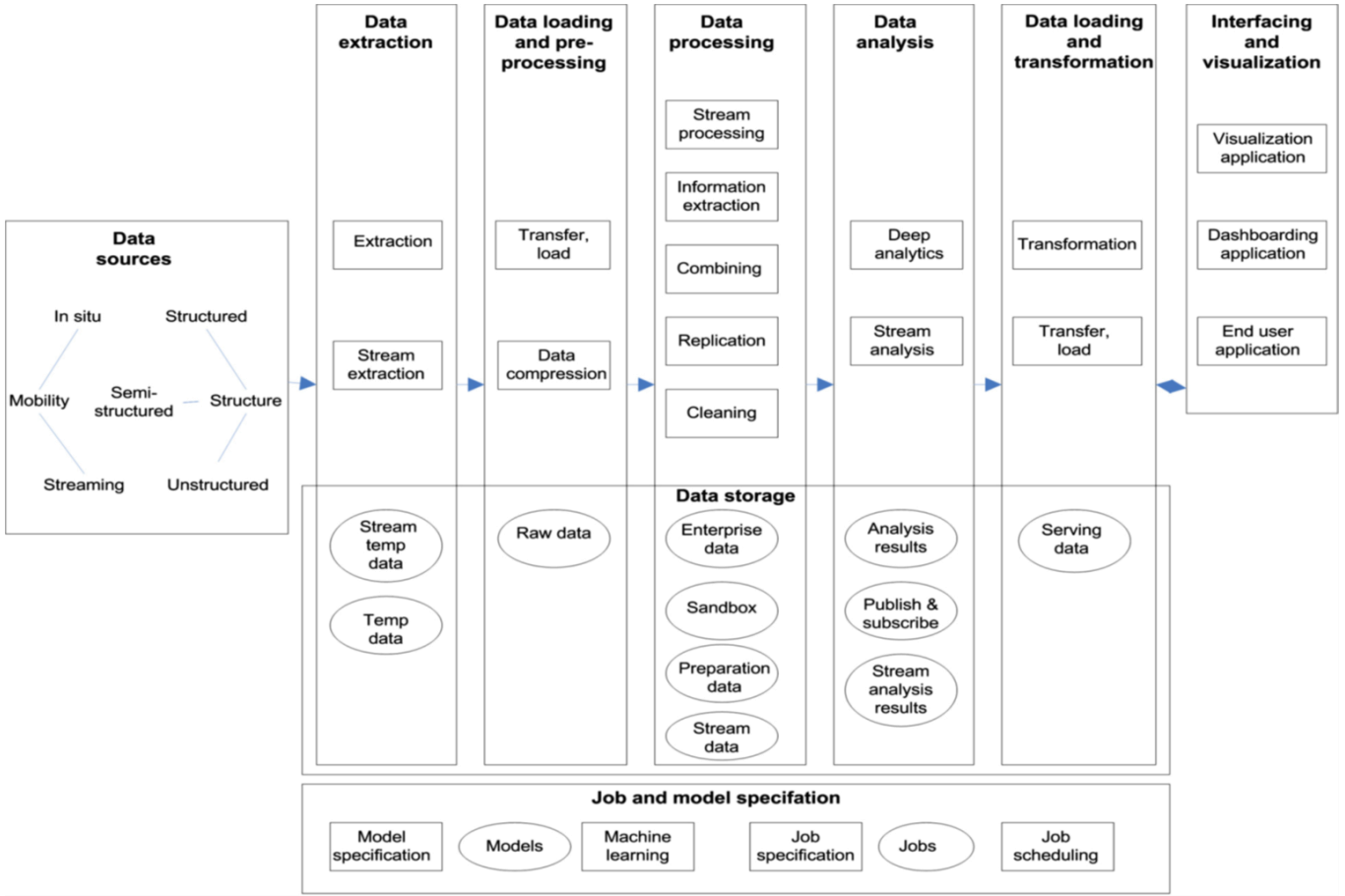
Diego Alberto Rincón Yáñez MCSc.
Twitter: @d1egoprog.

Taxonomías de Arquitecturas de Big Data

¿Quiénes usan arquitecturas de referencia de big data?

Diego Alberto Rincón Yáñez MCSc.
Twitter: @d1egoprog.

Taxonomía de las Arq. de referencia de Big Data

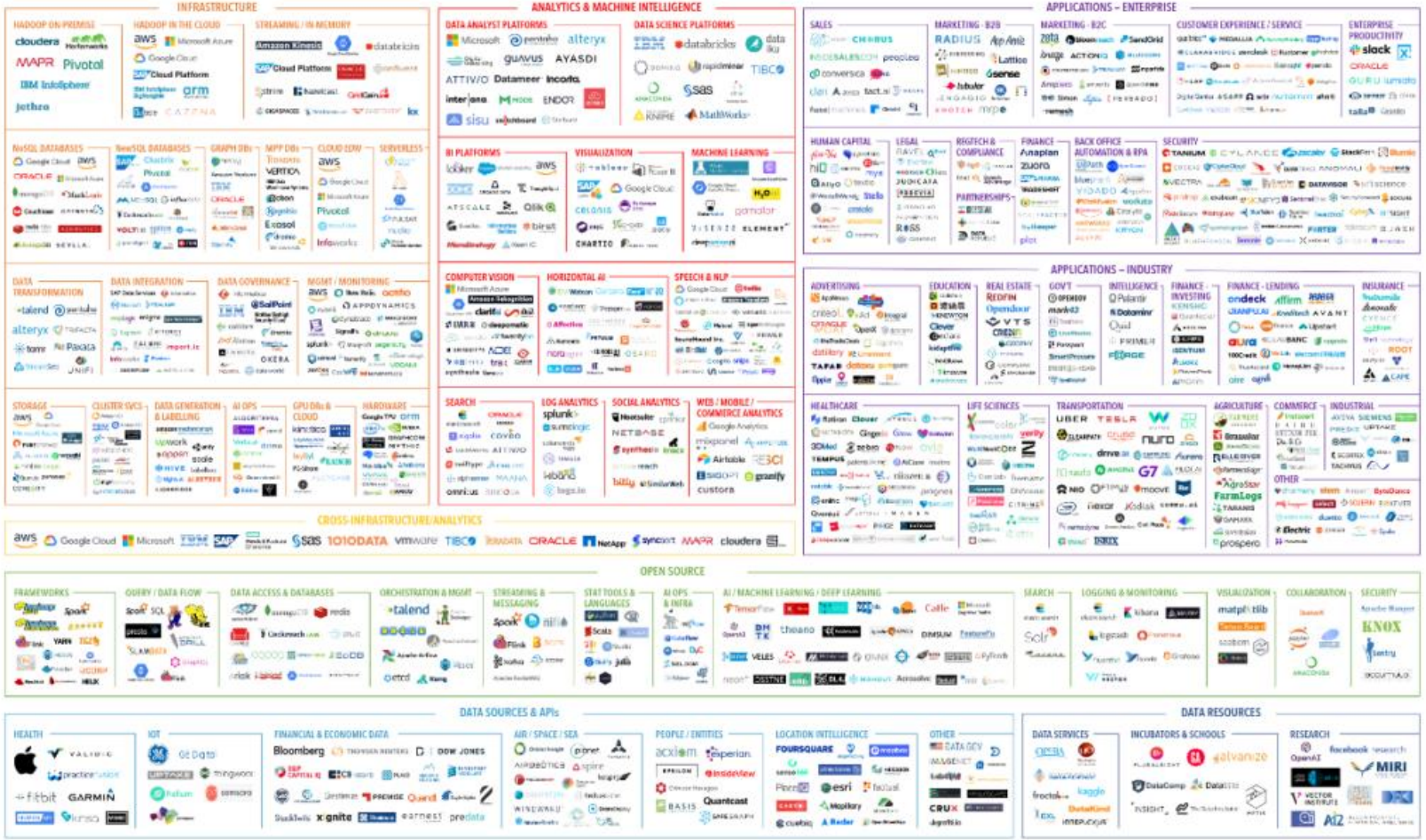


Diego Alberto Rincón Yáñez MCSc.
Twitter: @d1egoprog.

Tecnologías de Big Data

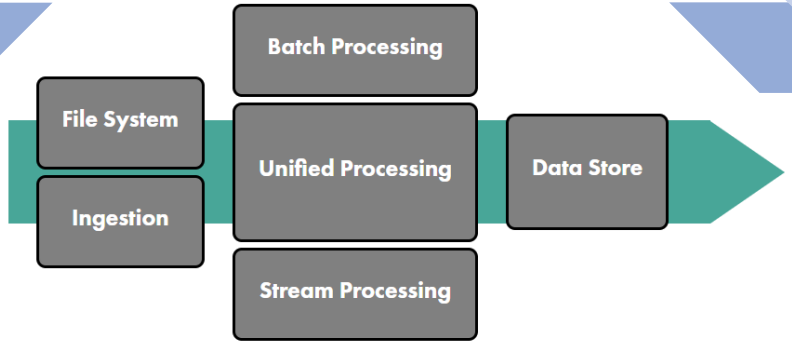
Diego Alberto Rincón Yáñez MCSc.
Twitter: @d1egoprog.

DATA & AI LANDSCAPE 2019



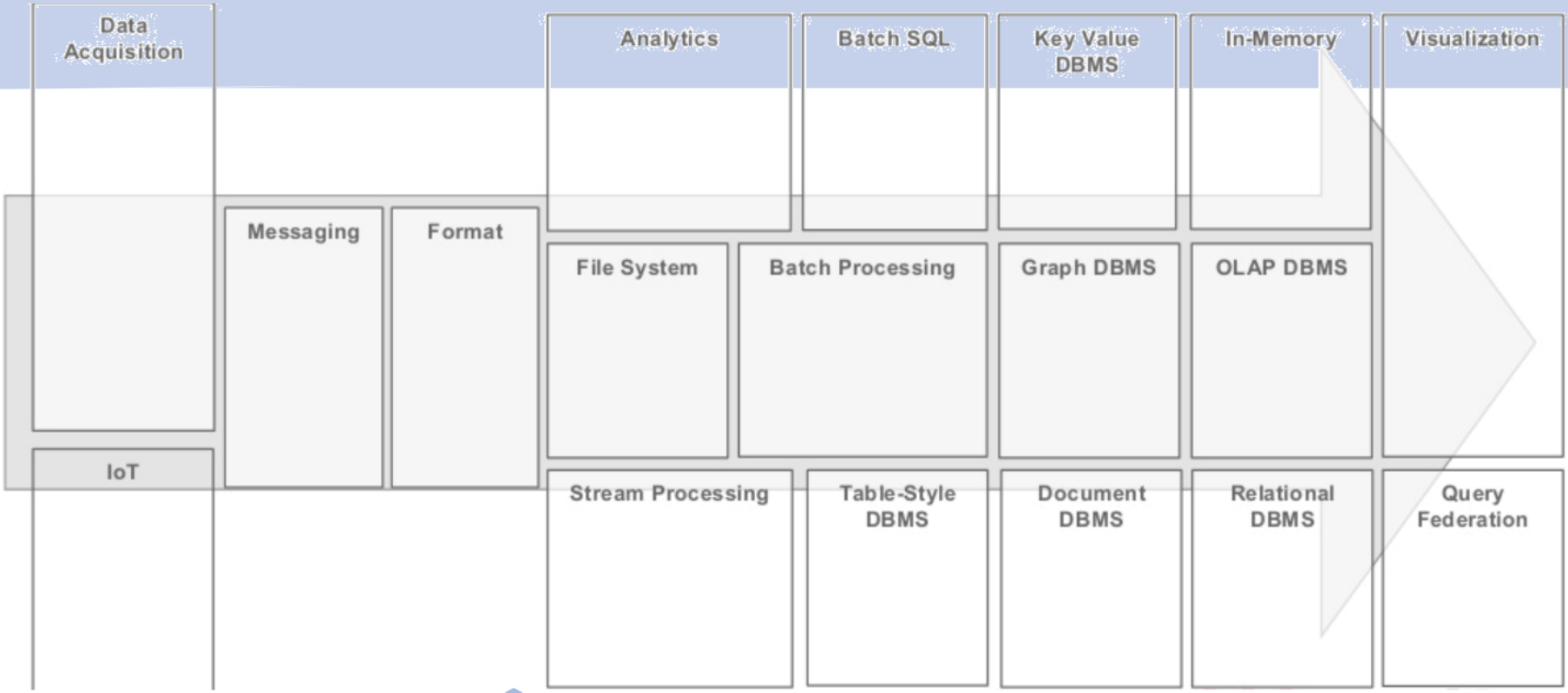
mattturck.com/data2019

FIRSTMARK
EARLY STAGE VENTURE CAPITAL



Diego Alberto Rincón Yáñez MCSc.
Twitter: @d1egoprog.

Plantilla de Tecnologías



Diego Alberto Rincón Yáñez MCSc.
Twitter: @d1egoprog.

Exploradas en este módulo

- Processing
  - YARN
  - MapReduce
- Analytics
  - SPARK
- NoSQL Database
  - MongoDB

Diego Alberto Rincón Yáñez MCSc.
Twitter: @d1egoprog.

Definición de proyecto

Diego Alberto Rincón Yáñez MCSc.
Twitter: @d1egoprog.

Descripción del Problema

Dentro de las problemáticas actuales, **buscar una fuente de información** continua que permita poner a prueba los conceptos utilizados en la materia de HPC, esta fuente deberá tener al menos **características** de **acceso a información continua, heterogeneidad en los datos, generación de datos constante**.

Ejemplos:

- Monitoreo de Tráfico en Internet o Redes Públicas
- Rutina implementando un spider en webs específicas
- Repositorios de datos (de cualquier tipo)
- Logs Generados por cualquier servicio (interno o externo)

Para lo anterior, deberán utilizar las arquitecturas, herramientas, conceptos vistos en clase y en la maestría para, construir una **pequeña plataforma** enfocada en el análisis información en **tiempo real** y **posible predicción** indicadores según el dominio del problema.

Diego Alberto Rincón Yáñez MCSc.
Twitter: @d1egoprog.

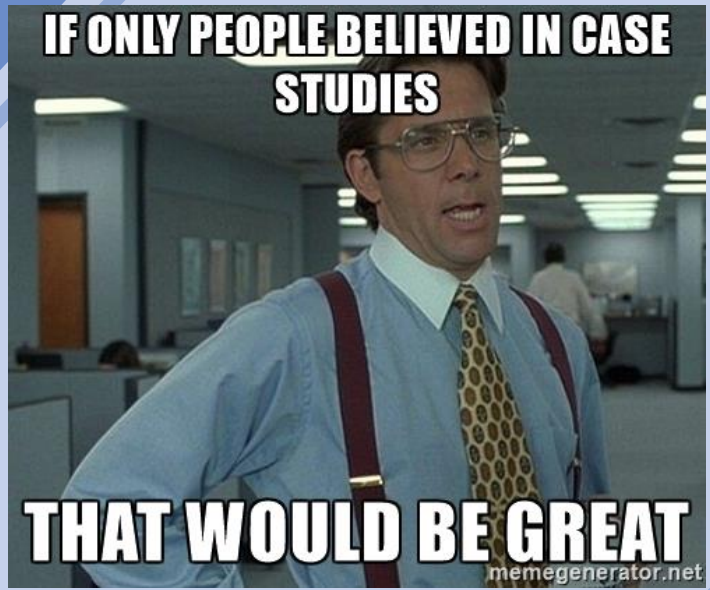
Reglas de Entrega

- Aval del Problema (Hilo en Blackboard para ese fin)
- Usar máquinas asignadas.
- Sustentación final 21 de Noviembre
- Nota por todo el equipo (Con aportes personales).
- Mínimos de Entrega
  - Debe tener instalado y hacer uso de:
    - ✓ Spark
    - ✓ MongoDB en HA.
    - ✓ HDFS + YARN + MapReduce.
  - Documentación del proyecto (Formatos en Blackboard)
    - Documentación de(los) datasets (ingesta de datos).
    - Documentación de Arquitectura.
    - Flujo de los datos y transformaciones en Información.

Calendario de Entregas

- Entrega Parcial 1 (Jueves 7 de Noviembre)
  - Presentación (30 minutos)
    - Alcance de Solución
    - Descripción de alcance analítico (# Indicadores).
    - Definición inicial de Arquitectura
    - Blueprint de tecnologías
- Entrega Final (Jueves 21 de Noviembre)
  - Documentación Completa
  - Prototipo Funcional + Montaje de Servicios
  - Presentación de resultados (10 Minutos)

Diego Alberto Rincón Yáñez MCSc.
Twitter: @d1egoprog.



Casos de Estudio

- 5 Grupos – 3G4P y 2G3P
- Creación de Hilo de discusión en Blackboard
  - Aval de Presentación

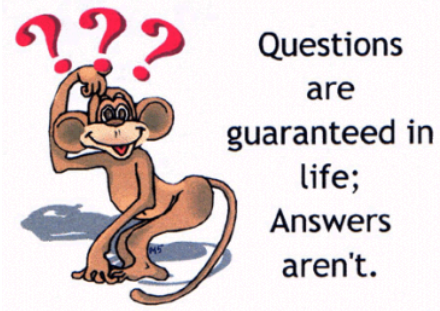
Diego Alberto Rincón Yáñez MCSc.
Twitter: @d1egoprog.

Reglas de Presentación

- Introducción de la Problemática
- Contexto de la Solución
  - Funcionalidades
  - Alcance
  - Estado Actual de la Solución (Bibliografía nueva??)
- Explicación de la Arquitectura
  - Restricciones
  - Ventajas
- Flujo de datos y transformaciones.
- Blueprint de Tecnologías
  - Ecosistema e integración.
  - Descripción y funcionalidad (por Tecnología)
  - Entradas/Salidas (por Tecnología)
- Aprendizajes y Factores a Resaltar
- Conclusiones

31 de Octubre (20 Minutos)

Diego Alberto Rincón Yáñez MCSc.
Twitter: @d1egoprog.



¿Preguntas?

Diego Alberto Rincón Yáñez MCSc.
Twitter: @d1egoprog.