

# Conceptos Fundamentales de de Big Data

Computación de Alto Desempeño  
Diego Alberto Rincón Yáñez MCSc  
[diego-rincon@javeriana.edu.co](mailto: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 MSc.  
Twitter: @d1egoprog.

# Antes de empezar.....

## ¿Qué es Big Data?



Diego Alberto Rincón Yáñez MSc.  
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 MSc.  
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.



# ¿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



# Agenda

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



Diego Alberto Rincón Yáñez MSc.  
Twitter: @d1egoprog.



# Big Data & Data Science



Diego Alberto Rincón Yáñez MSc.  
Twitter: @d1egoprogram

# Balanceo de Términos

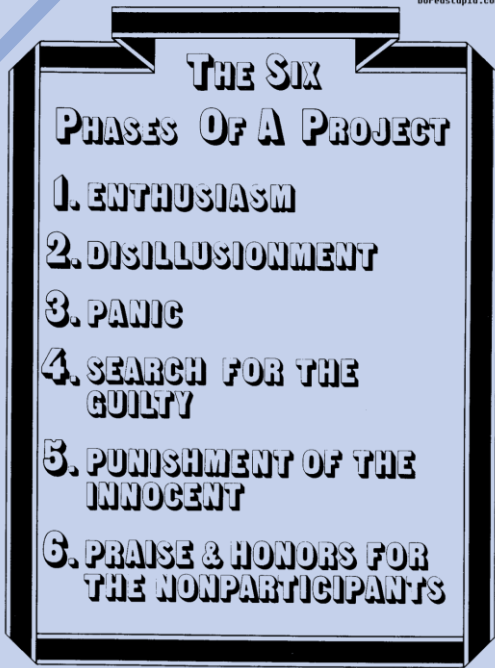
- Analítica (Des, Diag, Pred y Pres)
- Data Science vs Big Data
- Arquitectura (Sw, IT, BD, Empr)
- B.D. (Architech, Engineer, Scientist)
- CDO (Chief Data Officer)
- Business Intelligence
- KPI (Key Performance Indicators)
- Open (Data vs Access)
- Metadata
- ETL vs ELT
- OLTP vs OLAP
- Data (Warehouse,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



# 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.



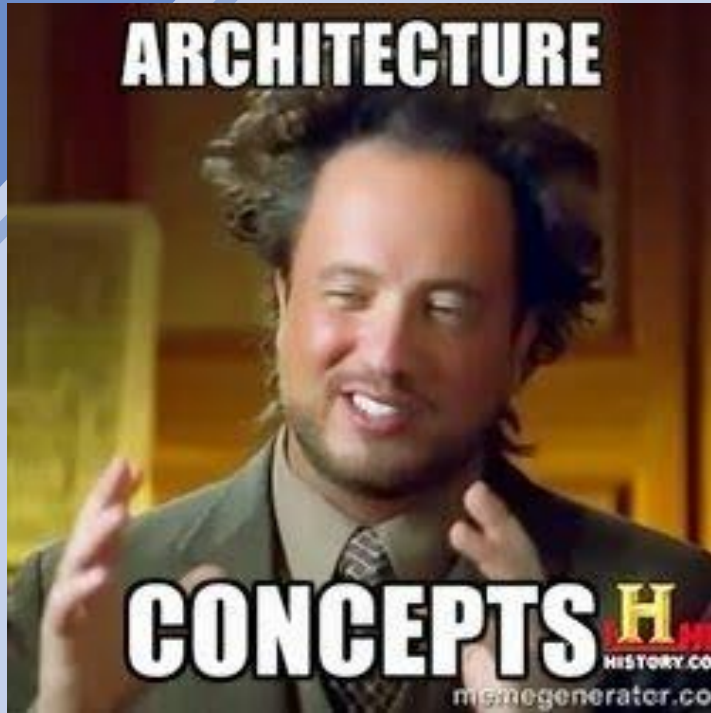


# Fases del proceso



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**



## Current Favorite Product List

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



## Current Product Count

4

**Information => derived**



Diego Alberto Rincón Yáñez MSc.  
Twitter: @d1egoprog.

# Operacional

## Características

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





# 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 MSc.  
Twitter: @d1egoprog.

# Arquitecturas en Números

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



# 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.

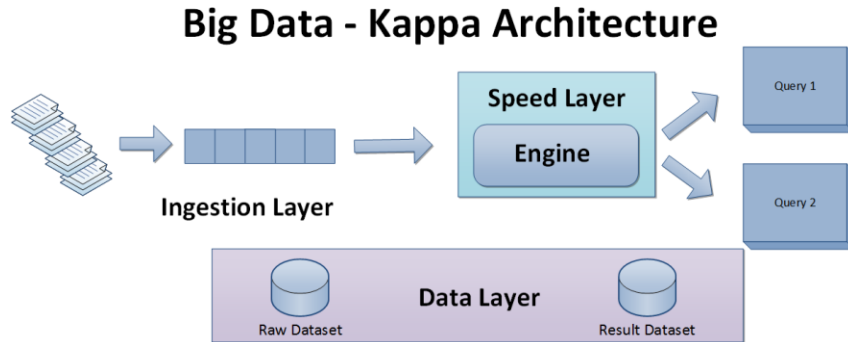


# Arquitecturas estándares para Big Data



Diego Alberto Rincón Yáñez MSc.  
Twitter: @d1egoprog.

# Arquitectura Kappa

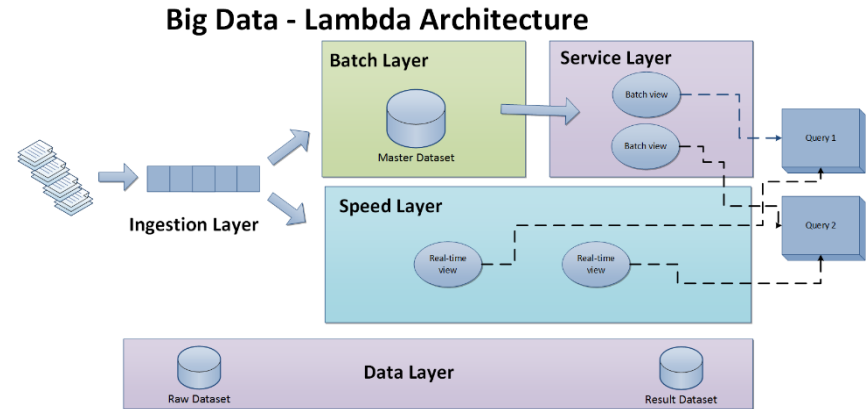


- la capa de datos
- la capa de velocidad



# 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

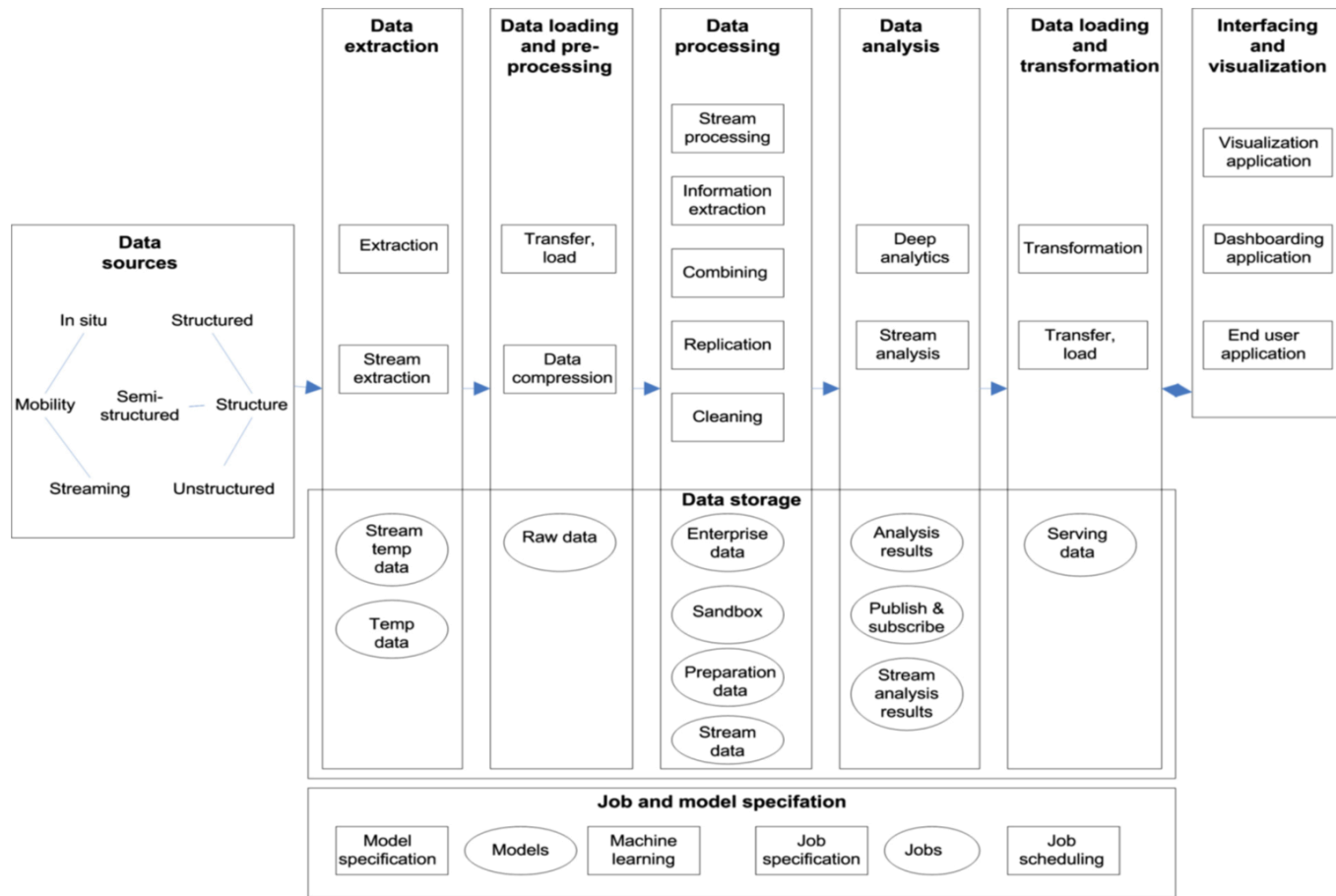


¿Quiénes usan arquitecturas de referencia de big data?



Diego Alberto Rincón Yáñez MCSc.  
Twitter: @d1egoprogram

# Taxonomía de las Arq. de referencia de Big Data



Diego Alberto Rincón Yáñez MCSc.  
Twitter: @d1egoprog.



# Tecnologías de Big Data



Diego Alberto Rincón Yáñez MSc.  
Twitter: @d1egoprog.

## DATA &amp; AI LANDSCAPE 2019

## INFRASTRUCTURE

The collage features logos for various cloud and data technologies, organized into three main sections:

- HADOOP ON-PREMISE:** cloudstack, MapR, Pivotal, IBM InfoSphere, jethro.
- HADOOP IN THE CLOUD:** AWS, Google Cloud, Microsoft Azure, IBM Cloud Platform, IBM Cloud.
- STREAMING / IN MEMORY:** Amazon Kinesis, Databricks, Amazon EMR, Cloudera, and others.

[illegible]

**DATA TRANSFORMATION**

- Alteryx
- Tableau
- Informatica
- Qlik
- Microsoft
- Google
- Amazon
- IBM
- SAP
- Oracle
- Microsoft
- Google
- Amazon
- IBM
- SAP
- Oracle

**DATA INTEGRATION**

- Informatica
- Microsoft
- Google
- Amazon
- IBM
- SAP
- Oracle
- Microsoft
- Google
- Amazon
- IBM
- SAP
- Oracle

**DATA GOVERNANCE**

- Informatica
- Microsoft
- Google
- Amazon
- IBM
- SAP
- Oracle
- Microsoft
- Google
- Amazon
- IBM
- SAP
- Oracle

**MGMT./ARCHITECTURE**

- Informatica
- Microsoft
- Google
- Amazon
- IBM
- SAP
- Oracle
- Microsoft
- Google
- Amazon
- IBM
- SAP
- Oracle

**STORAGE**

- IBM
- Microsoft Azure
- Google Cloud
- Amazon S3
- Oracle
- NetScout Systems
- NetScout Systems
- NetScout Systems

**CLUSTER GCES**

- IBM
- Microsoft Azure
- Google Cloud
- Amazon EC2
- Oracle
- NetScout Systems
- NetScout Systems
- NetScout Systems

**DATA GENERATION & LABELLING**

- IBM
- Microsoft Azure
- Google Cloud
- Amazon EC2
- Oracle
- NetScout Systems
- NetScout Systems
- NetScout Systems

**AI OPS**

- IBM
- Microsoft Azure
- Google Cloud
- Amazon EC2
- Oracle
- NetScout Systems
- NetScout Systems
- NetScout Systems

**GPU DaaS & CLOUD**

- IBM
- Microsoft Azure
- Google Cloud
- Amazon EC2
- Oracle
- NetScout Systems
- NetScout Systems
- NetScout Systems

**HYPERSCALE**

- IBM
- Microsoft Azure
- Google Cloud
- Amazon EC2
- Oracle
- NetScout Systems
- NetScout Systems
- NetScout Systems



## ANALYTICS &amp; MACHINE INTELLIGENCE

[illegible]

The collage is organized into three main sections, each with a title in red capital letters:

- BI PLATFORMS:** Includes logos for Looker, Amazon Redshift, AWS, Tableau, Microsoft Power BI, SAP, Google Cloud, Alteryx, Qlik, and IBM.
- VISUALIZATION:** Includes logos for CRANIS, Tableau, Microsoft Power BI, SAP, Google Cloud, Alteryx, Qlik, and IBM.
- MACHINE LEARNING:** Includes logos for Amazon SageMaker, Microsoft Azure Machine Learning, Google Cloud AI, H2O, and Element.

**COMPUTER VISION**

- Microsoft Azure
- Clarifai
- IBM Watson
- Amazon Rekognition
- Google Cloud Vision
- Facebook AI
- OpenAI
- DeepMind
- Scale AI
- Imaginary
- Scale AI

**HORIZONTAL AI**

- IBM Watson
- Google Cloud
- Amazon Rekognition
- Facebook AI
- OpenAI
- DeepMind
- Scale AI
- Imaginary
- Scale AI

**SPEECH & NLP**

- Google Cloud
- Amazon Rekognition
- Facebook AI
- OpenAI
- DeepMind
- Scale AI
- Imaginary
- Scale AI

The collage displays logos for the following tools:

- SEARCH:** Searchmetrics, SEMrush, Ahrefs, Moz, Majestic, Omniture, SiteCrawler.
- LOG ANALYTICS:** Splunk, Sumologic, Loggly, Logstash, Elastic, Logz.io.
- SOCIAL ANALYTICS:** Hootsuite, Sprinklr, Netbase, Brandwatch, Social Searcher, Bitly, SimilarWeb.
- WEB / MOBILE / COMMERCE ANALYTICS:** Google Analytics, Mixpanel, Amplitude, Airbrake, ESI, SiftOpt, and Grizzly.

## OPEN SOURCE

The banner displays a wide array of open-source software logos, organized into 12 distinct categories. The categories and their associated logos are as follows:

- FRAMEWORKS:** Includes logos for Spring, Vaadin, and others.
- QUERY / DATA FLOW:** Includes logos for Apache Spark, Databricks, and others.
- DATA ACCESS & DATABASES:** Includes logos for Apache Cassandra, Apache HBase, and others.
- ORCHESTRATION & INGEST:** Includes logos for Apache Airflow, Apache NiFi, and others.
- STREAMING & MESSAGING:** Includes logos for Apache Kafka, Apache Pulsar, and others.
- START TOOLS & LANGUAGES:** Includes logos for Docker, Kubernetes, and others.
- AI OPS / INFRA:** Includes logos for Argo CD, Jenkins, and others.
- AI / MACHINE LEARNING / DEEP LEARNING:** Includes logos for TensorFlow, PyTorch, and others.
- SEARCH:** Includes logos for Elasticsearch, Solr, and others.
- LOGGING & MONITORING:** Includes logos for Prometheus, Grafana, and others.
- VISUALIZATION:** Includes logos for Tableau, Power BI, and others.
- COLLABORATION:** Includes logos for Slack, Microsoft Teams, and others.
- SECURITY:** Includes logos for Apache Wicket, Apache Shiro, and others.

#### DATA SOURCES & APIs

A horizontal banner displaying logos for various companies, organized into several categories:

- HEALTH**: Includes logos for Apple, Veeva Systems, Paracore Labs, Fitbit, Garmin, and Juniper.
- IoT**: Includes logos for GE Data, IBM Watson, and others.
- FINANCIAL & ECONOMIC DATA**: Includes logos for Bloomberg, Thomson Reuters, Dow Jones, Capital IQ, Moody's, S&P Global, Standard & Poor's, Morningstar, and others.
- AIR / SPACE / SEA**: Includes logos for Airbus, Boeing, SpaceX, and others.
- PEOPLE / ENTITIES**: Includes logos for LinkedIn, Experian, and others.
- LOCATION INTELLIGENCE**: Includes logos for Foursquare, Mapbox, Esri, and others.
- OTHER**: Includes logos for Data City, Amazon Web Services, and others.

## APPLICATIONS – ENTERPRISE

[illegible]

## APPLICATIONS – INDUSTRY

[illegible]

## DATA RESOURCES

**DATA SERVICES**

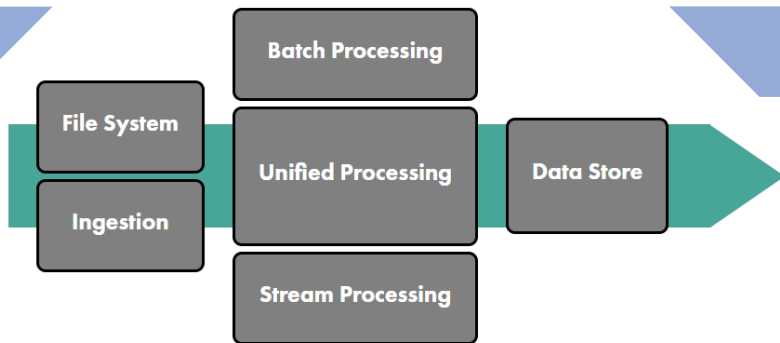
- datacamp
- dataiku
- fractal
- lms
- dataiku
- dataiku

**INCUBATORS & SCHOOLS**

- FLORALAB
- DataCamp
- DataCamp
- INCUBATOR
- INCUBATOR

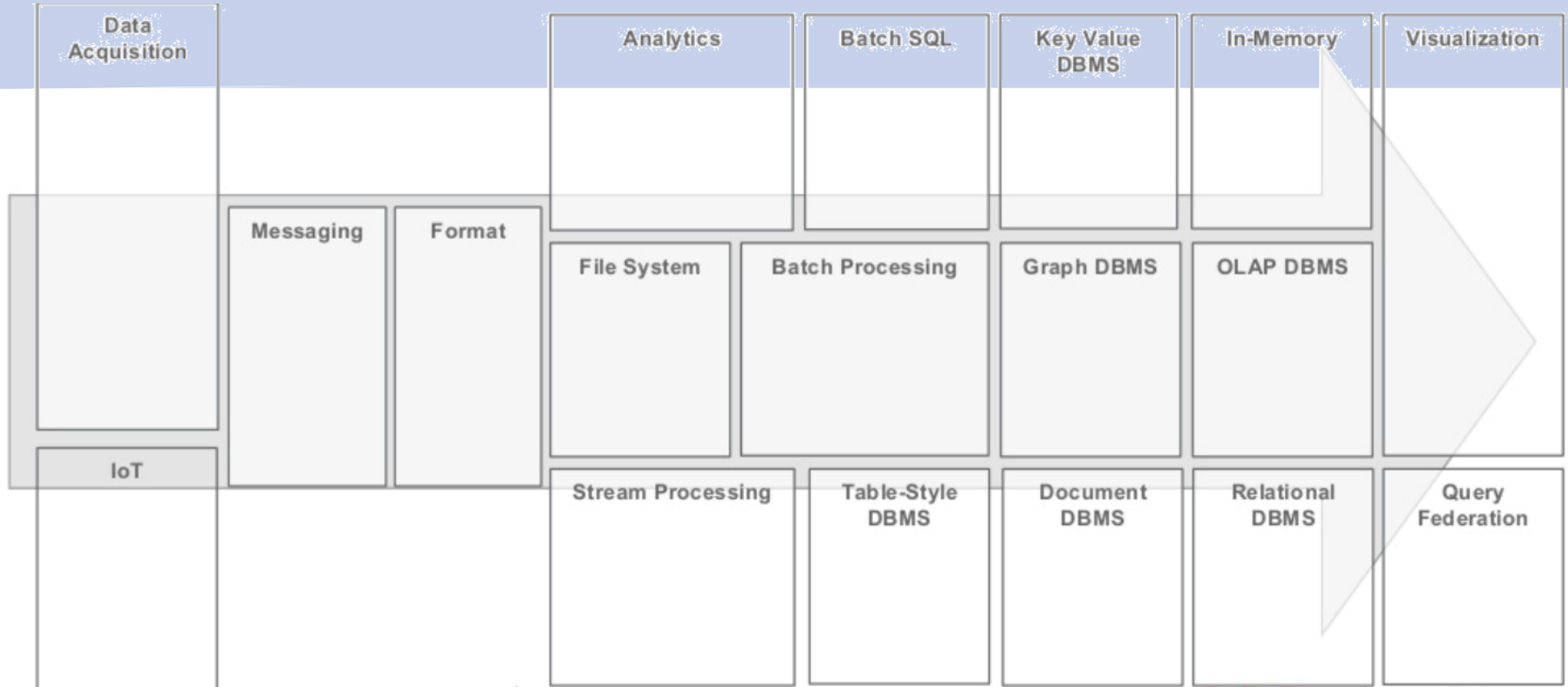
**RESEARCH**

- OpenAI
- facebook research
- MIRI
- VECTOR INSTITUTE
- AI2
- AI2



Diego Alberto Rincón Yáñez MSc.  
Twitter: @d1egoprog.

# Plantilla de Tecnologías



Diego Alberto Rincón Yáñez MSc.  
Twitter: @d1egoprog.

# Exploradas en este módulo

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



# Definición de proyecto



Diego Alberto Rincón Yáñez MSc.  
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 MSc.  
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.



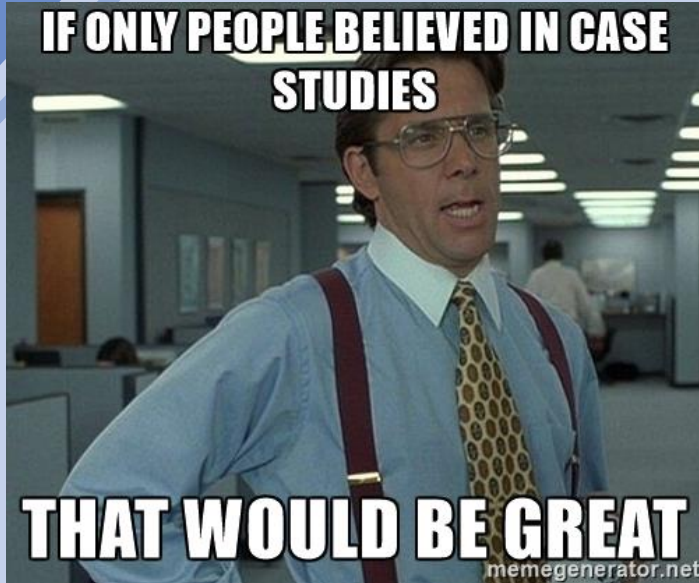
# Calendario de Entregas

- 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)



# Casos de Estudio

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



Diego Alberto Rinc  n Y  nez MSc.  
Twitter: @d1egoprogram

# 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 MSc.  
Twitter: @d1egoprog.



Questions  
are  
guaranteed in  
life;  
Answers  
aren't.

¿Preguntas?



Diego Alberto Rincón Yáñez MCSc.  
Twitter: @d1egoprog.