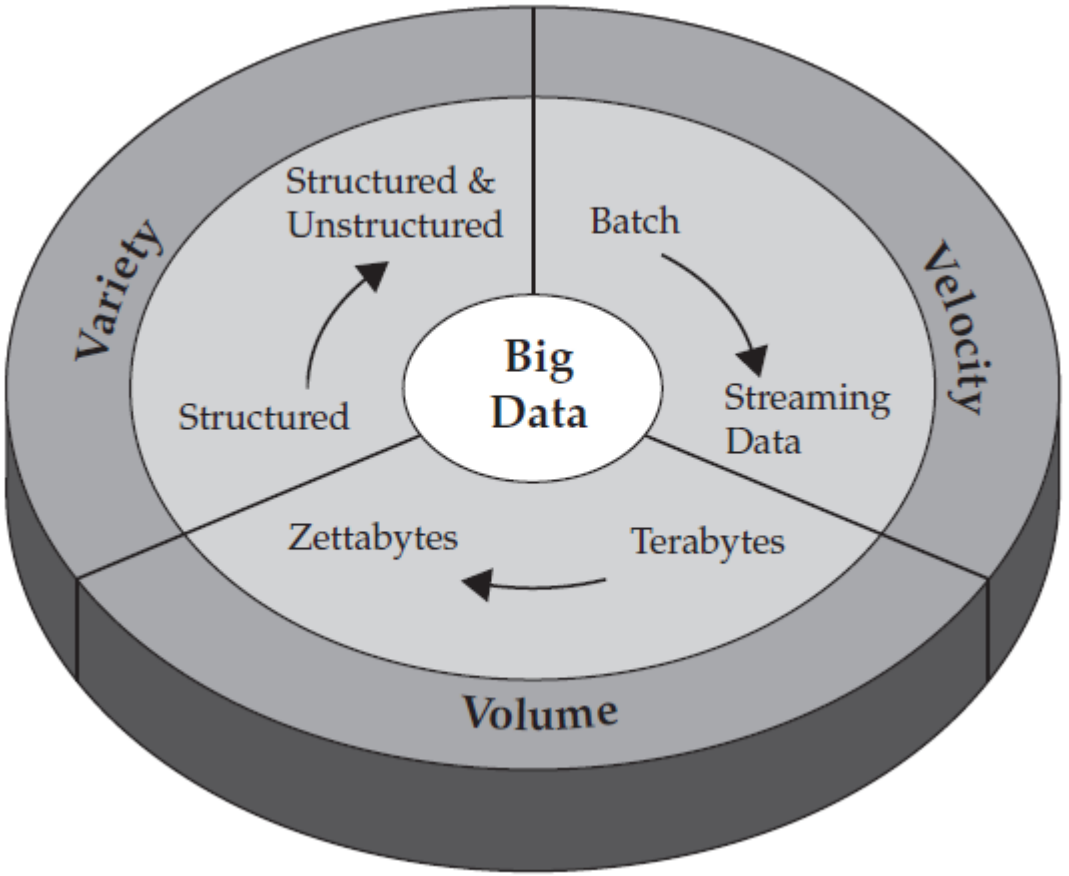
Big Data

Alejandro Sierra

2019

¿Qué es Big Data?



Herramientas

- Bases de Datos NoSQL
  - MongoDB, Cassandra, Hbase, ….
- Herramientas de Infexación y Recuperación de Información
  - Elasticsearch, Solr, Lucene, …
- Clusters de Procesamiento en Paralelo y Data Lakes
  - Hadoop, HDFS, Spark,
- Herramientas de Streaming
  - Kafka, NiFi, SparkStreaming, …

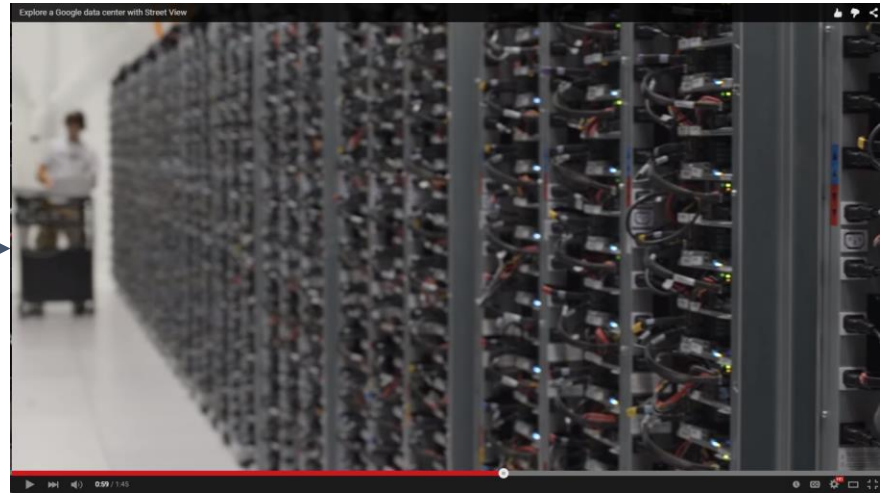
Big Data – Desafios técnicos

Escalabilidad Horizontal : Crecer en cantidad de nodos (máquinas)

- -Volumen
- -Paralelizar
- -Comunicación
- -Sincronización
- -Balancear carga
- -Manejo de fallos

Big Data - Motivación

Data Centers



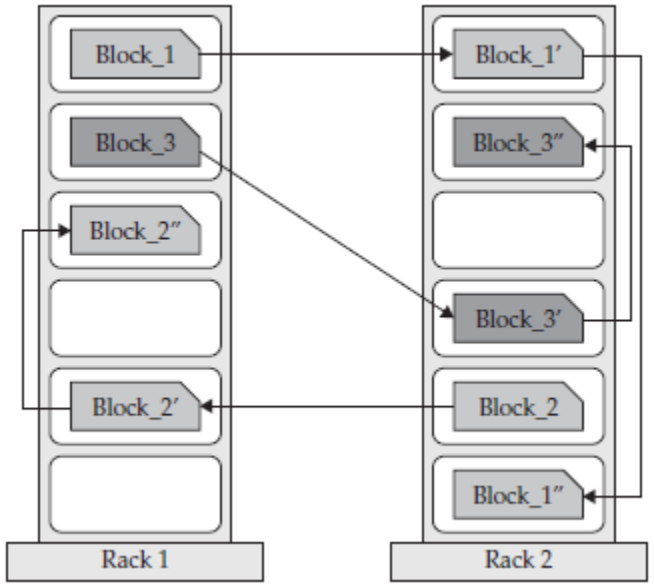
Hadoop

- Ideas base Google (MapReduce - GFS)
- Implementación de Apache
  - Yahoo!, Facebook, Twitter, LinkedIn, …
- MapReduce (Google) – Hadoop Map Reduce
  - Paralelismo, Cercania con los datos, Balance de carga
- GFS (Google File System) - HDFS
  - Particionar los datos
  - Replicación

HDFS -GFS

- Los datos se dividen en *bloques* de **64** MB (por defecto)
- Los bloques se copian (por defecto **2** copias extras) en diferentes nodos.
  - Tolerancia a Fallos
  - Alta disponibilidad
- La idea es tener los datos cerca a su procesamiento.
  - NAS o SAN no son convenientes en este caso.
  - *Data locality*. Cercanía de datos con el procesamiento.

HDFS

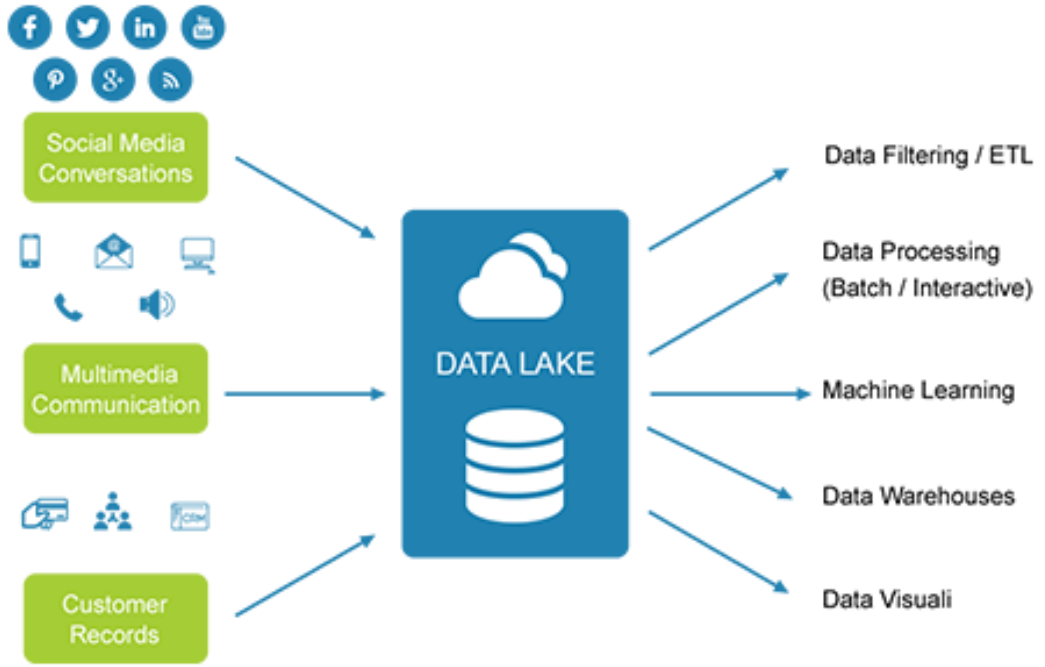


Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data
Dirk deRoos, Chris Eaton, George Lapis, Paul Zikopoulos, Tom Deutsch

Data Lake

- Repositorio de datos crudos
- Estructurado y no estructurados
- Sin esquema

- “Schema on read”
  - ELT



http://www.elevondata.com/services/data-lake

Madurez de Data Lakes

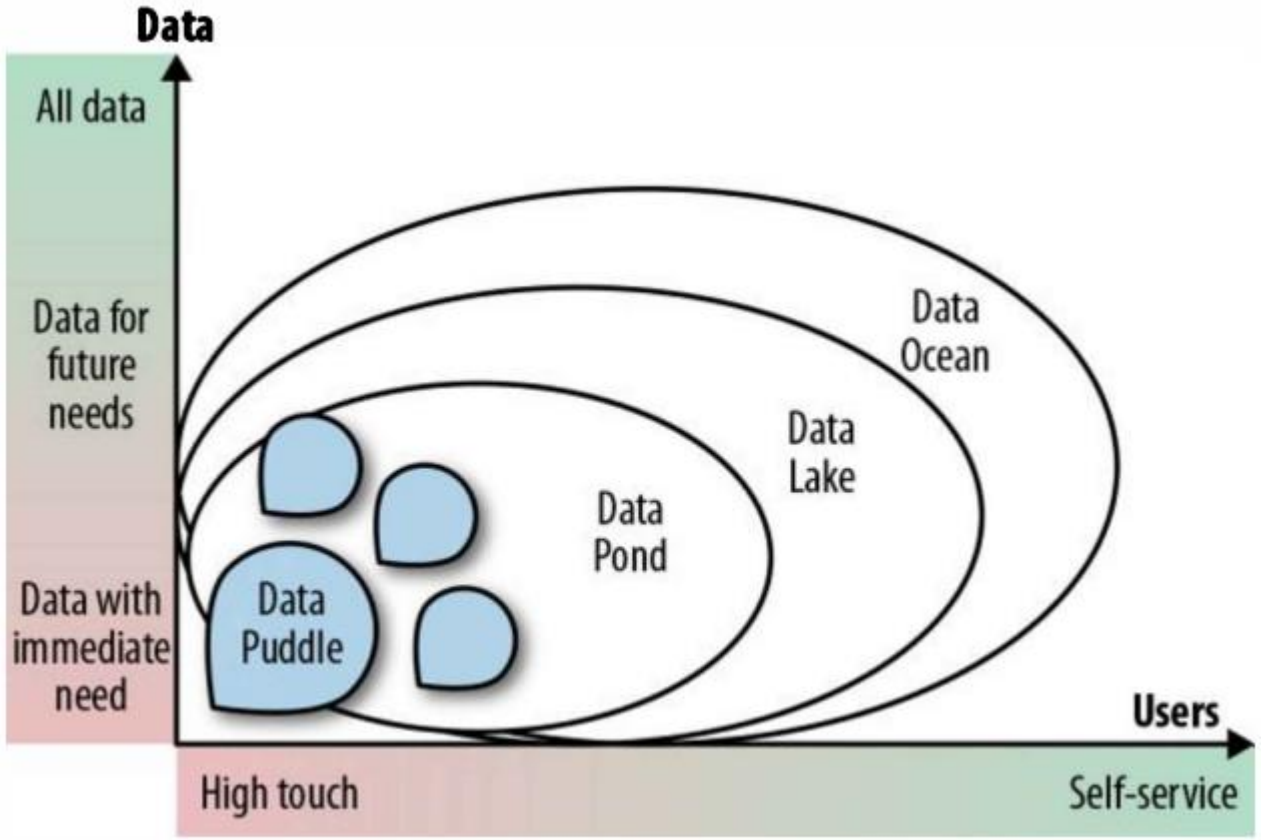


Figure 1-1. The four stages of maturity

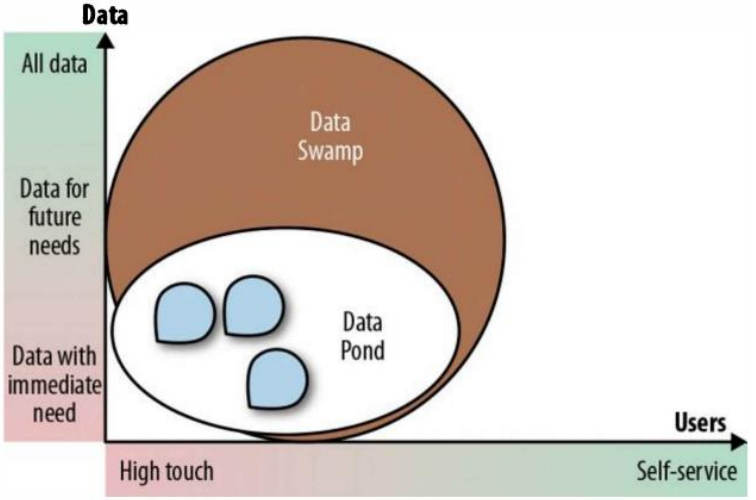


Figure 1-6. A data swamp

The enterprise big data lake delivering the promise of big data and data science / Alex Gorelik. Sebastopol, California iO'Reilly Media, Inc., 2019.

Zonas en un Data Lake

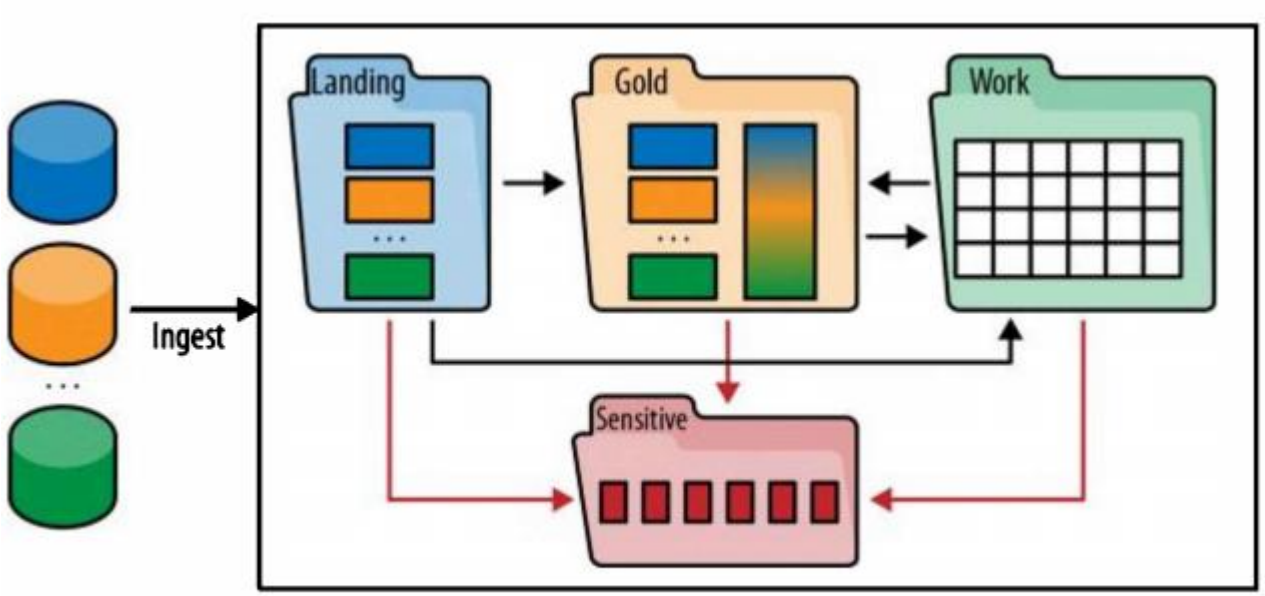


Figure 1-8. Zones of a typical data lake

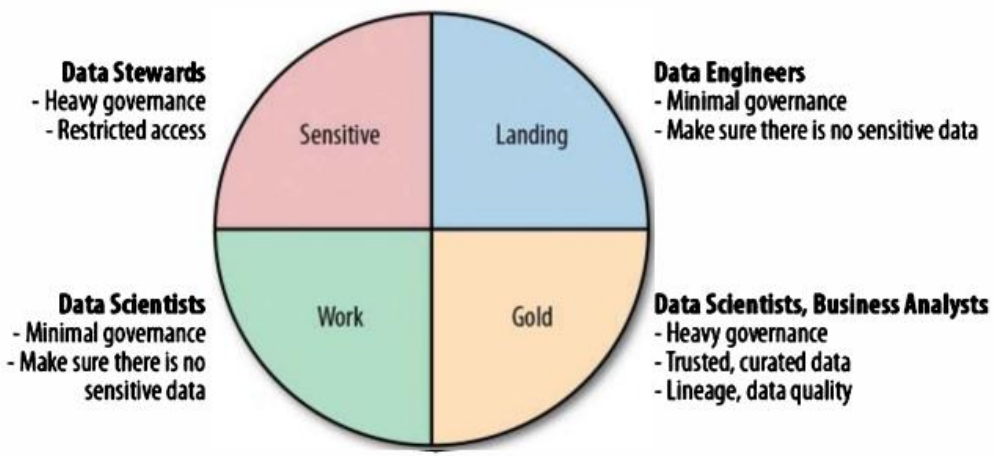


Figure 1-9. Governance expectations, zone by zone

MapReduce

- Paradigma de programación. (Programación funcional)
- Diseñado para ser ejecutado en paralelo en clusters grandes
  - Equipos sencillos que se pueden encontrar en el mercado.
- Se ocupa de abstraer los detalles técnicos y simplificar el procesamiento en 2 funciones **Map** y **Reduce**
- Single Instruction, Multiple Data (SIMD)

MapReduce - Ejemplo

Contador de palabras:

Tenemos una gran cantidad de textos y queremos saber cuales son las palabras más comunes en nuestra colección.

Map y Reduce

El programador solo debe especificar 2 funciones

map (in_key, in_value) -> list(out_key, intermediate_value)

- Recibe una pareja *llave , valor*.
- Emite un conjunto de nuevas parejas.
- Las llaves de entrada no son necesariamente las llaves de salida.

reduce (out_key, list(intermediate_value)) -> list(out_value)

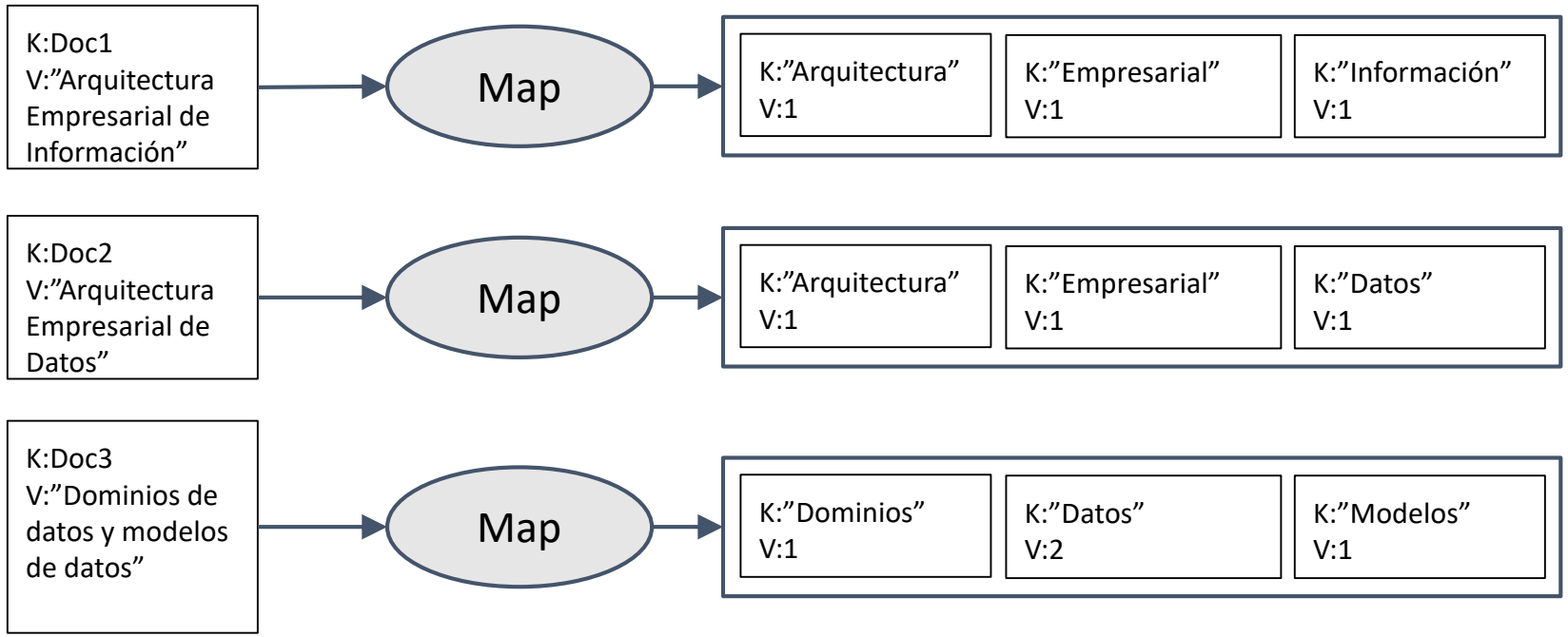
- Recibe una *llave* de las emitidas por **map** acompañada de la lista de *valores* generados en las diferentes instancias de **map**.
- Emite los valores finales
- Similar a GROUP BY (SQL)

MapReduce - Ejemplo

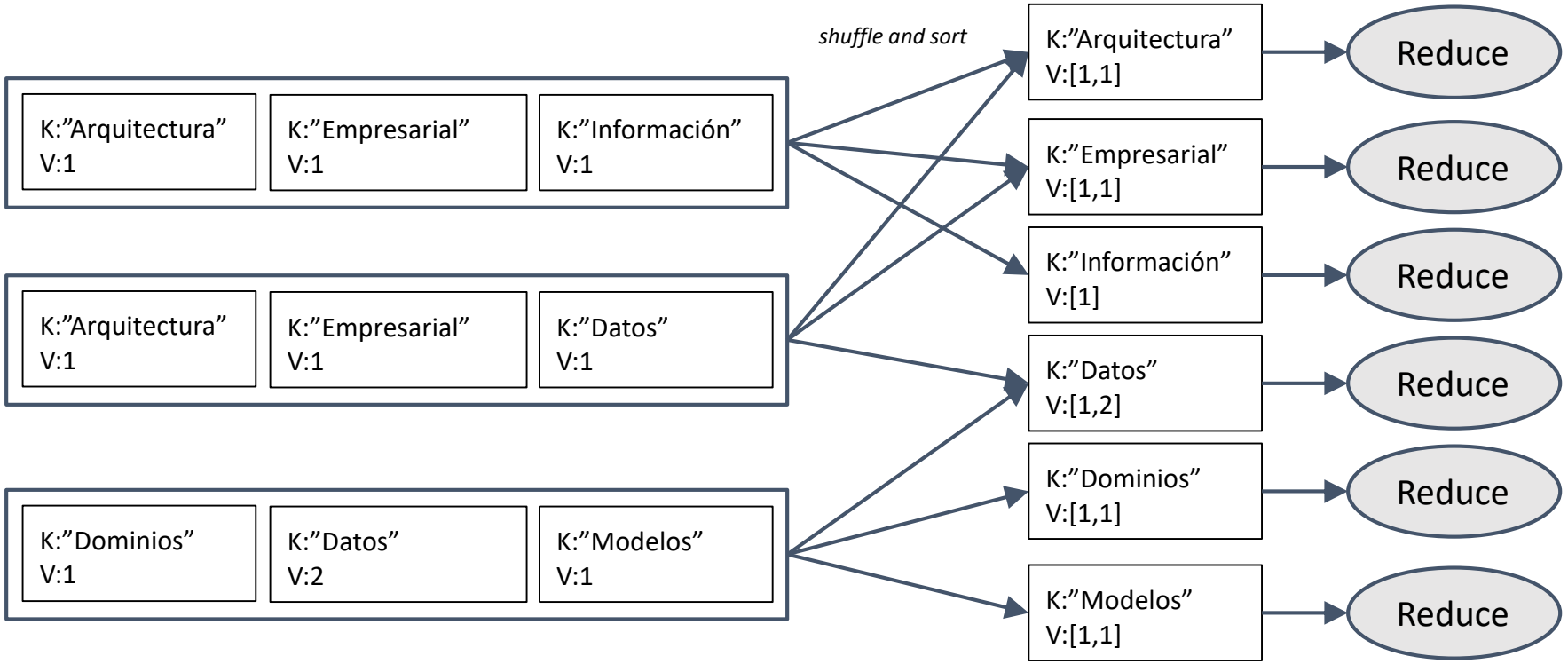
```
map(String input_key, String input_value):
        // input_key: document name
        // input_value: document contents
        for each word w in input_value:
                EmitIntermediate(w, "1");


reduce(String output_key, Iterator intermediate_values):
        // output_key: a word
        // output_values: a list of counts
        int result = 0;
        for each v in intermediate_values:
                result += ParseInt(v);
        Emit(AsString(result));
```

Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: simplified data processing on large clusters. Commun. ACM 51, 1 (January 2008), 107-113. DOI=10.1145/1327452.1327492
http://doi.acm.org/10.1145/1327452.1327492
http://research.google.com/archive/mapreduce.html

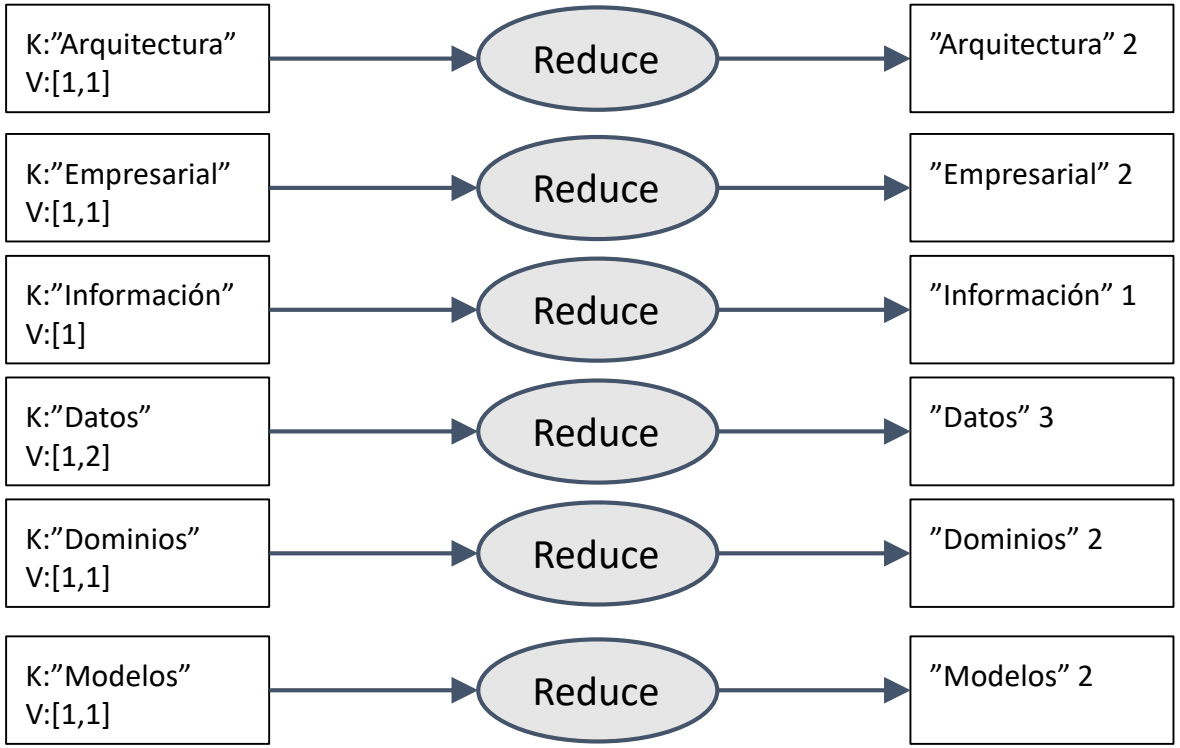
MapReduce - Ejemplo Map



MapReduce - Ejemplo



MapReduce - Ej. Reduce



Apache Hive

- Datos crudos (HDFS) -> HiveQL (Similar a SQL)
  - Reunir
  - Agrupar
  - Seleccionar
  - … datos en los archivos crudos
- Internamente
  - MapReduce/Tez/Spark

Datos vs Metadatos

- Metadatos describen Datos
- Metadatos describen la manera de acceder/interpretar los Datos

Nombre	Documento de Identificación	Oficina Registro	Saldo
Juan Pérez	65789157	Carrera 13 # 65-20	$ 2567897
Luis López	8741987	Calle 100 # 11-15	$ 4569787
María Rodríguez	97456812	Cr 3 # 10-09	$ 6547893
Claudia Gómez	52147896	Cr 68 # 13-47	$ 1597534

Metadatos

Datos

Hive external tables

- HDFS : /tmp/tweets_staging2/*.json

  Datos

- Hive:

```
CREATE EXTERNAL TABLE tweets (
        id_str STRING,
        retweet_count INT
)
...
LOCATION ' /tmp/tweets_staging2/ ';
```

  Metadatos

id_str	retweet_count
898300779605479424	3
898300779605479483	0
898300779605479495	Null
898300779605479499	1

Crear tabla con datos en CSV

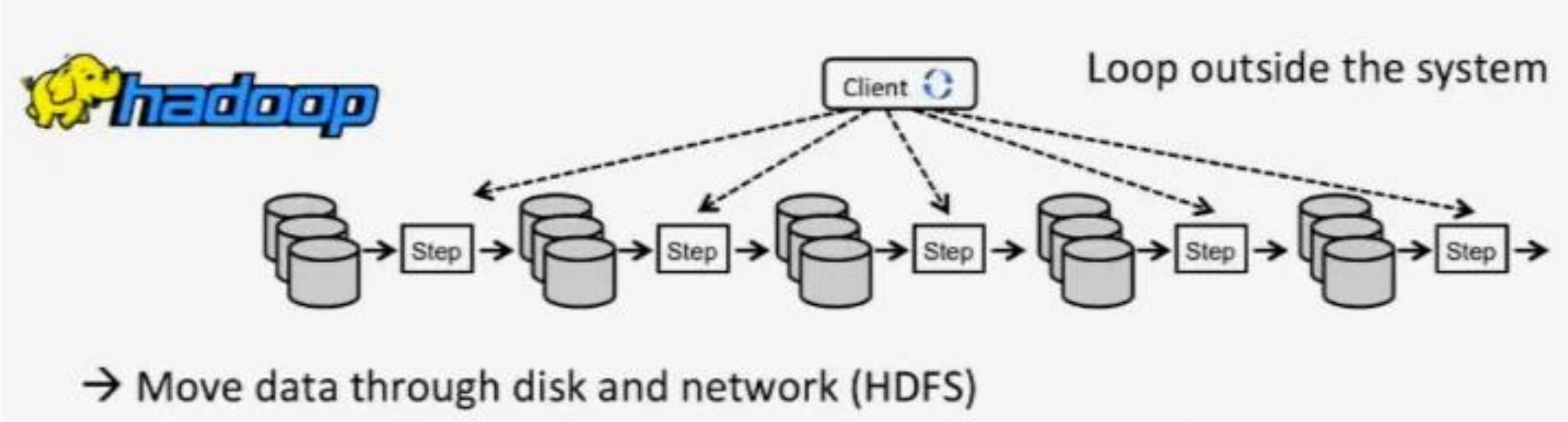
- SerDe
  - Serialize
  - Deserialize

- Luego de ejecutar la creación se pueden hacer consultas HiveQL

```
CREATE EXTERNAL TABLE Languages(
    SK_Language STRING,
    ISO639_3Code STRING,
    ISO639_2BCode STRING,
    ISO639_2TCode STRING,
    ISO639_1Code STRING,
    LanguageName STRING,
    Scope STRING,
    Type STRING,
    MacroLanguageISO639_3Code STRING,
    MacroLanguageName STRING,
    IsChild STRING
)
ROW FORMAT SERDE 'com.bizo.hive.serde.csv.CSVSerde'
WITH SERDEPROPERTIES (
   "separatorChar" = '\073',
   "quoteChar"     = "\"",
   "escapeChar"    = "\\"
)
stored as textfile
 LOCATION '/tmp/Languages'
 tblproperties ("skip.header.line.count"="1");
```

Hadoop MapReduce - Desventajas

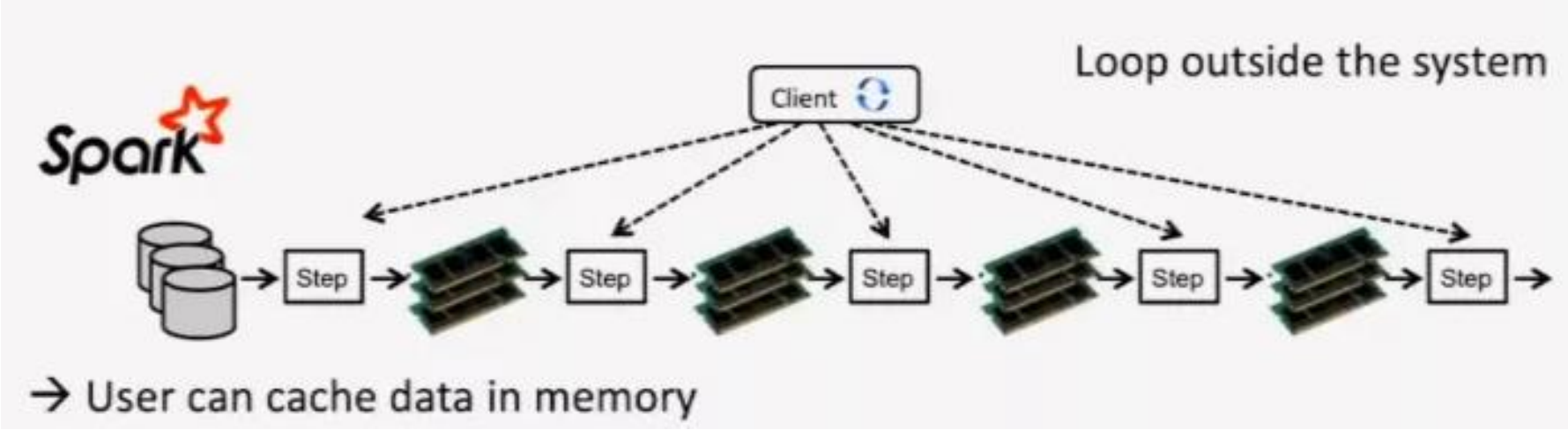
- Continuamente acceso a disco
- Map/Reduce puede no resolver todo tipo de problemas de manera fácil.



https://acadgild.com/blog/hadoop-vs-spark-best-big-data-frameworks/

Apache Spark

- Ejecución en paralelo en un cluster
- Todos los datos están en memoria
  - Mucho más rápido que MapReduce

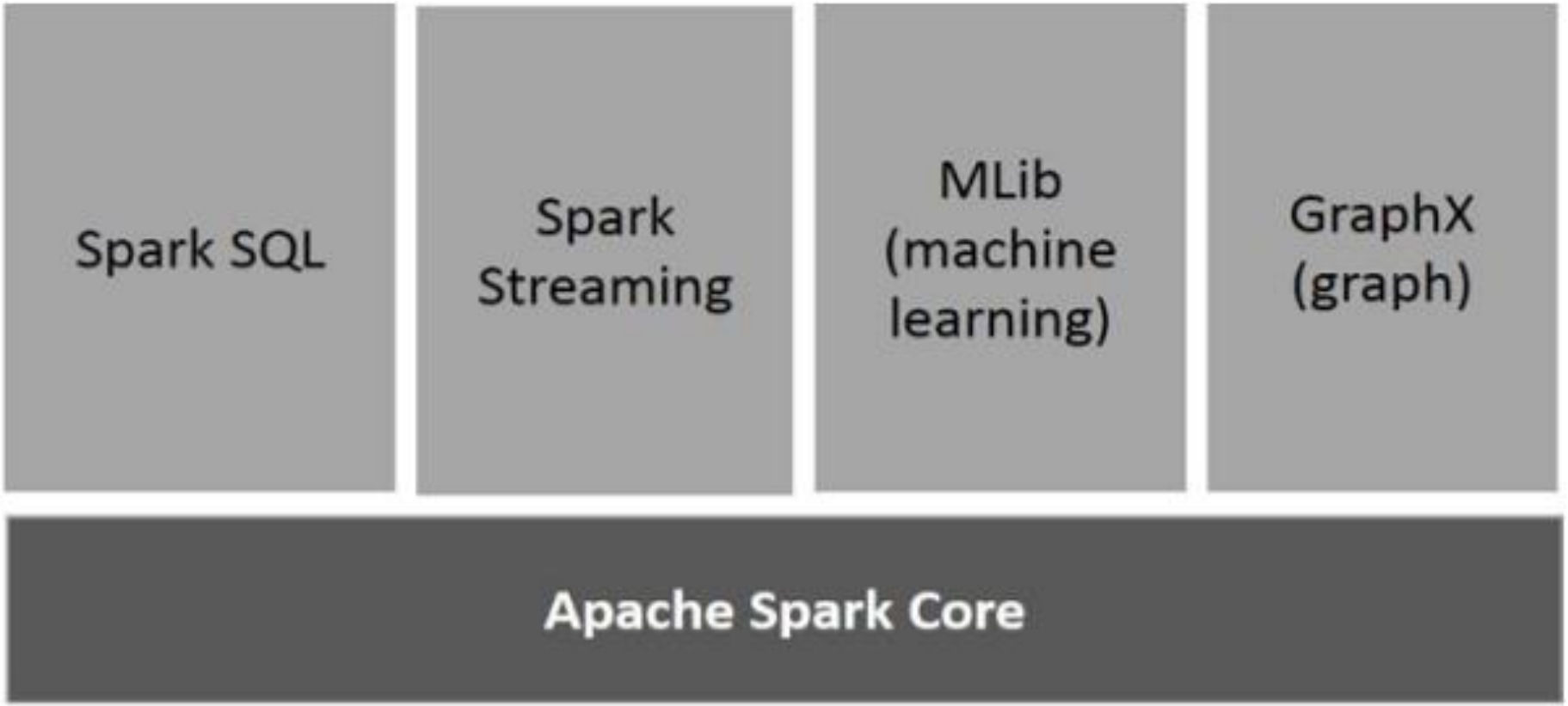


https://acadgild.com/blog/hadoop-vs-spark-best-big-data-frameworks/

No solo Map/Reduce

- Más funciones
  - Map
  - Filter
  - flatMap
  - Intersection
  - AggregateByKey
  - Reduce
  - Collect
  - Count
  - first

```
sparkContext.textFile("hdfs://...")
            .flatMap(line => line.split(" "))
            .map(word => (word, 1)).reduceByKey(_ + _)
            .saveAsTextFile("hdfs://...")
```



https://www.tutorialspoint.com/apache_spark/apache_spark_introduction.htm

Lenguajes

- Scala

- Java

- Python

- R

- SQL

SparkSQL

- Similar a Hive
- Convierte una consulta en SQL en una serie de pasos en Spark
- Puede consultar el catálogo de hive

- Definición de esquema

```
case class IrisSchema(sepalLength: Double, sepalWidth: Double,
petalLength: Double, petalWidth: Double, classIris: String)
```

- Relación datos-metadatos

```
val irisData = sc.textFile("/tmp/iris.data.csv").map(_.split(","
)).map(p => IrisSchema(p(0).toDouble, p(1).toDouble, p(2
).toDouble,p(3).toDouble,p(4))).toDF()


irisData.registerTempTable("irisData")
```

- Consultas SQL sobre la tabla resultante

```
select classIris,count(*) from irisData group by classIris
```

classIris	_c1
Iris-setosa	50
Iris-virginica	50
Iris-versicolor	50

Ecosistema Hadoop (Hortonworks)

https://es.hortonworks.com/ecosystems/

DWH vs Hadoop

DWH

- Datos estructurados
- Medidas bien establecidas
- Calidad – Limpieza
  - E.g. Sarbanes-Oxley
- Datos seleccionados
  - Valor percibido
- “Schema on write”

Data Lake

- Todo tipo de datos
- Descubrimiento de Información
- Datos crudos
- Todos los datos pueden tener valor
- “Schema on read”

SE COMPLEMENTAN

Responsabilidades de IT y Negoci

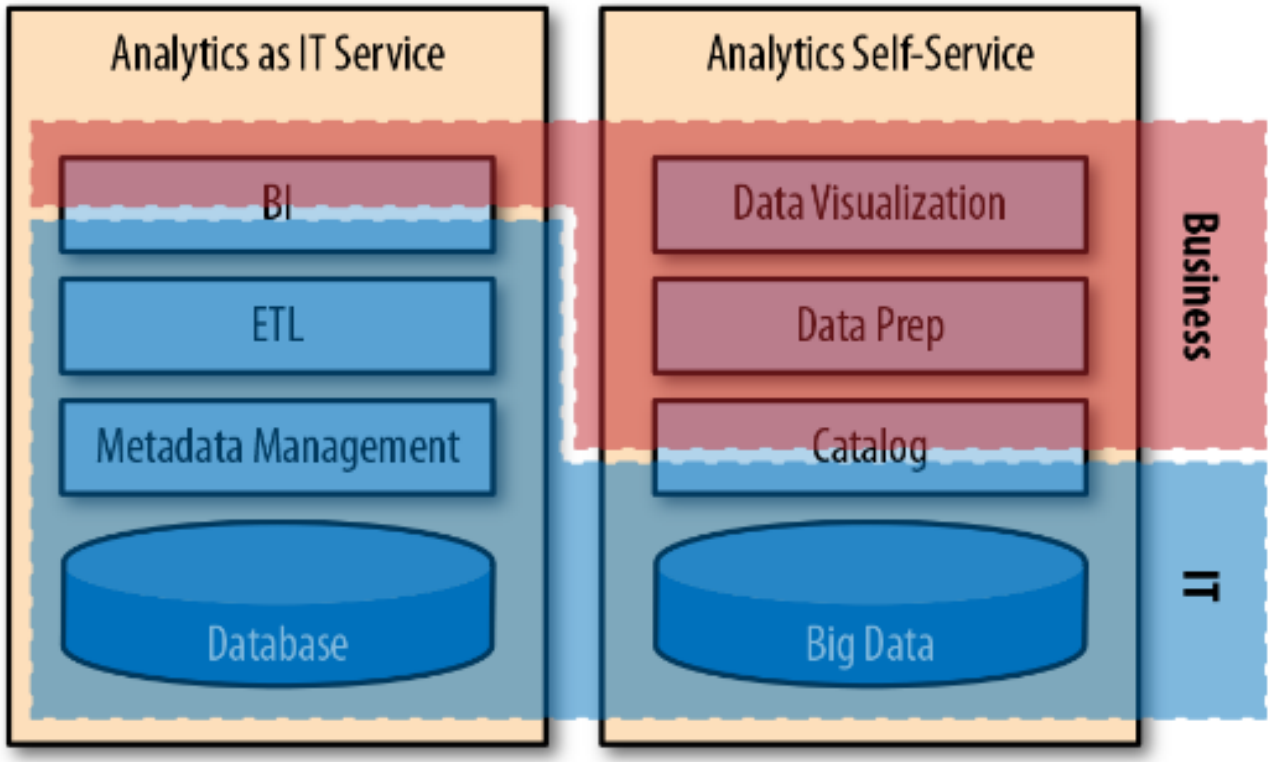


Figure 6-1. Enabling analysts and reducing the load on IT with self-service analytics

The enterprise big data lake delivering the promise of big data and data science / Alex Gorelik. Sebastopol, California iO'Reilly Media, Inc., 2019.

Referencias

- The enterprise big data lake delivering the promise of big data and data science / Alex Gorelik. Sebastopol, California iO'Reilly Media, Inc., 2019.
- Spark
  - https://spark.apache.org/
- Hadoop
  - http://hadoop.apache.org/
- Hortonworks
  - https://es.hortonworks.com/
- Cloudera
  - https://www.cloudera.com/
- Databricks
  - https://databricks.com/