PASS SQL saturday

#608 | BOGOTÁ 2017

Mayo 13, 2017

Bogotá, Colombia

#sqlsatBogota

# Patrocinadores del SQL Saturday
























# Conéctese con PASS

Regístrese hoy para una membresía gratis:

pass.org



#sqlpass

# Sea cual sea su pasión datos - hay un capítulo virtual para usted!

# AZURE DATA LAKE

Jorge Muchaypiña Gutierrez

Business Intelligence Specialist

MTA | MCP | MAP | MCSA | MCSE BI | ITILF | CSM

Blog: https://jorgemuchaypina.wordpress.com/

LinkedIn: https://pe.linkedin.com/in/jorge-muchaypiña-79038491

Facebook: https://www.facebook.com/jorgemichael.muchaypinagutierrez.5

PASS

PASS SQL saturday

#608 | BOGOTÁ 2017

# Que es un Big Data?

*"**Big Data** es como el sexo adolescente:*

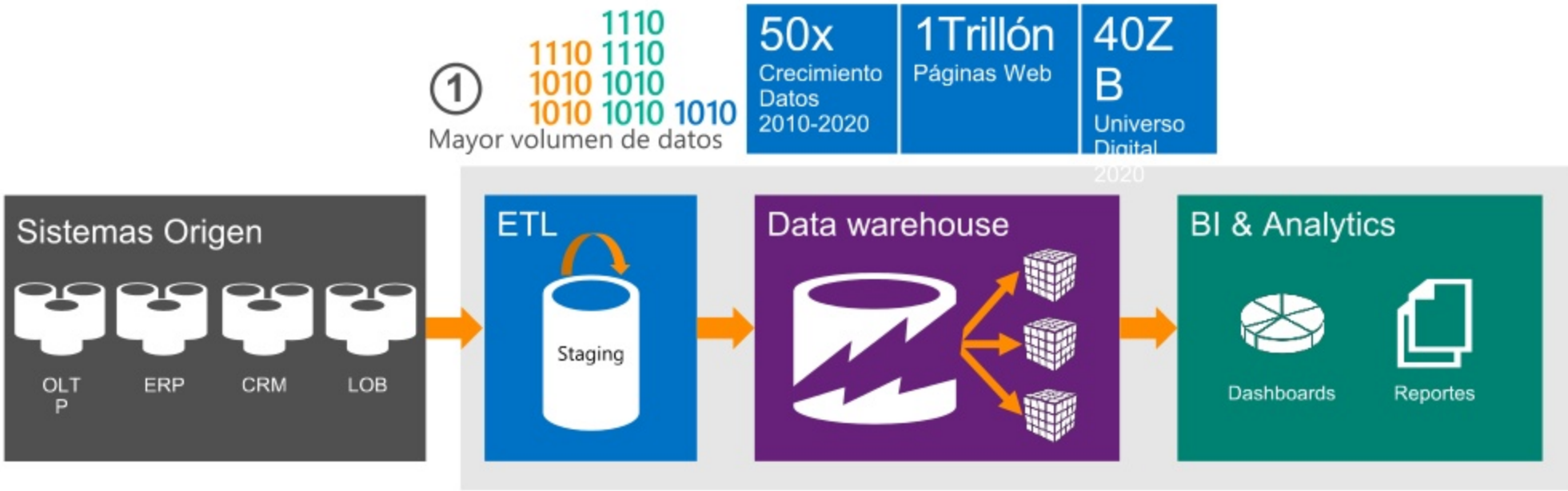
*todos hablan acerca de ello, nadie sabe realmente como hacerlo,*

*todos piensan que todos lo están haciendo,*

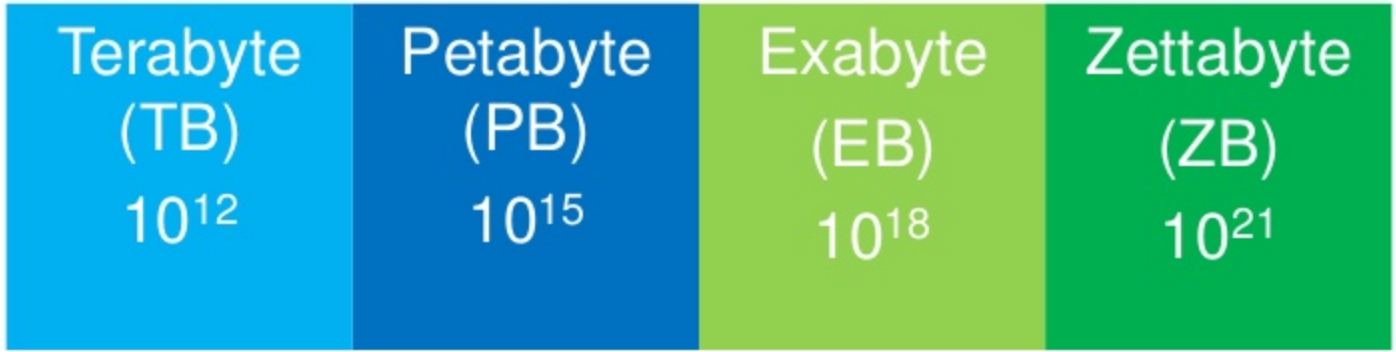
*por lo que todo el mundo dice que lo esta haciendo“*

*Dan Ariely*

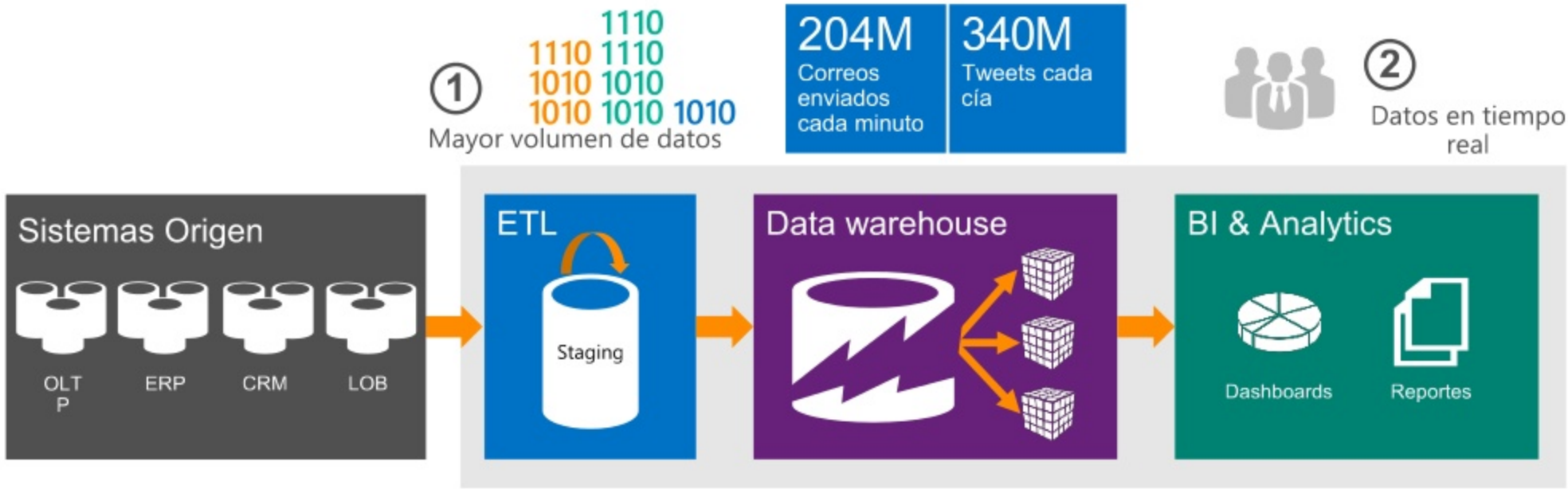
# Volumen


1110
1110 1110
1010 1010
1010 1010 1010
1
Mayor volumen de datos
50x
Crecimiento Datos 2010-2020
1Trillón
Páginas Web
40Z
B
Universo Digital
2020
Sistemas Origen
OLT P
ERP
CRM
LOB
ETL
Staging
Data warehouse
BI & Analytics
Dashboards
Reportes

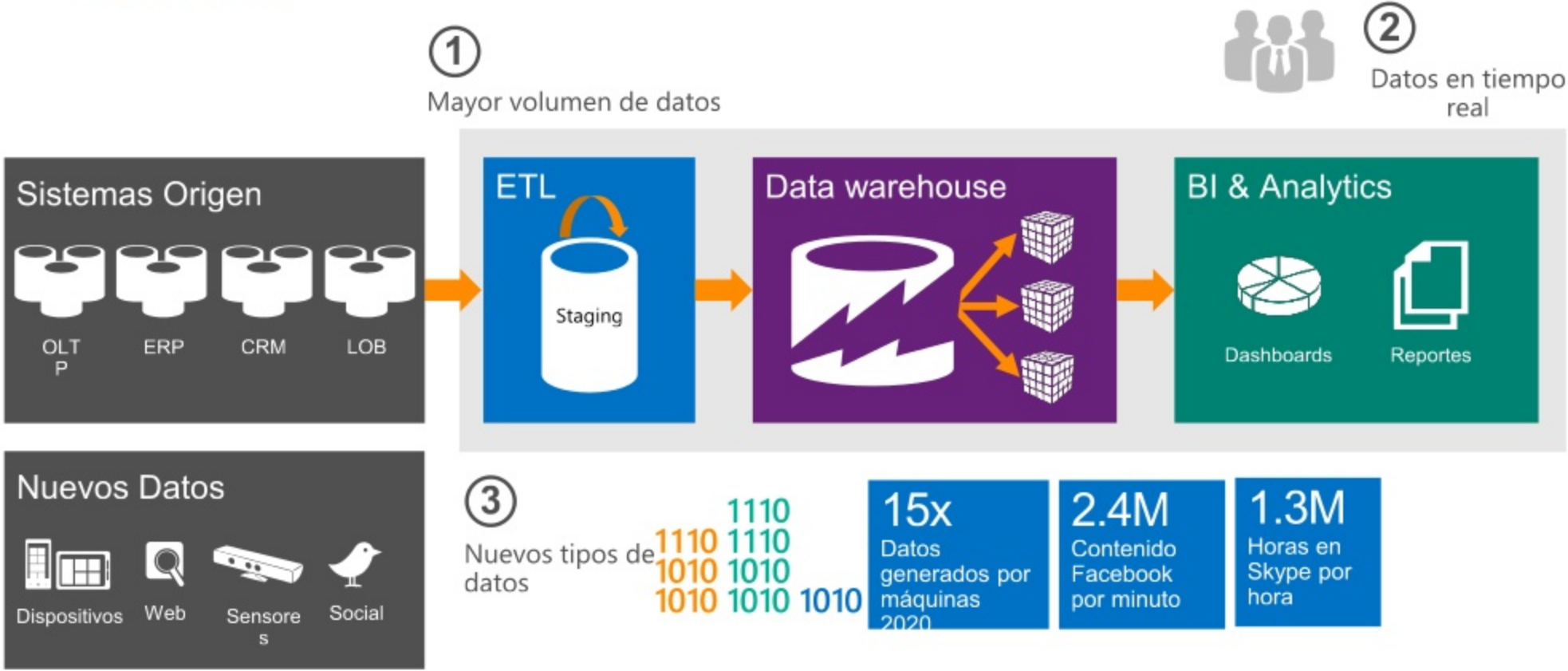
# Volumen


Terabyte
(TB)
$10^{12}$
Petabyte
(PB)
$10^{15}$
Exabyte
(EB)
$10^{18}$
Zettabyte
(ZB)
$10^{21}$

# Velocidad


1110
1110 1110
1010 1010
1010 1010 1010
1
Mayor volumen de datos
204M
Correos enviados cada minuto
340M
Tweets cada cía
2
Datos en tiempo real
Sistemas Origen
OLTP
ERP
CRM
LOB
ETL
Staging
Data warehouse
BI & Analytics
Dashboards
Reportes

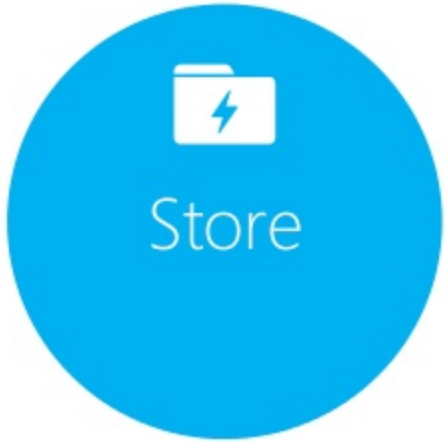
# Variedad


1
Mayor volumen de datos
2
Datos en tiempo real
Sistemas Origen
OLTP
ERP
CRM
LOB
ETL
Staging
Data warehouse
BI & Analytics
Dashboards
Reportes
Nuevos Datos
Dispositivos
Web
Sensores
Social
3
Nuevos tipos de datos
1110 1110 1010 1010 1010 1010 1010
15x
Datos generados por máquinas 2020
2.4M
Contenido Facebook por minuto
1.3M
Horas en Skype por hora

# The 3 Azure Data Lake Services


Clusters as a service


Big data queries as a service




Hyper-scale Storage optimized for analytics

# Apache Hadoop

## Hadoop almacena los archivos en un sistema de archivos distribuido

Almacenamiento y procesamiento distribuido entre múltiples servidores

Los archivos pueden estar en múltiples nodos

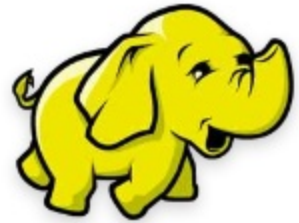
## Hadoop puede almacenar grandes volúmenes de información

El almacenamiento puede crecer con la demanda dependiendo del número de nodos

Escala de forma lineal

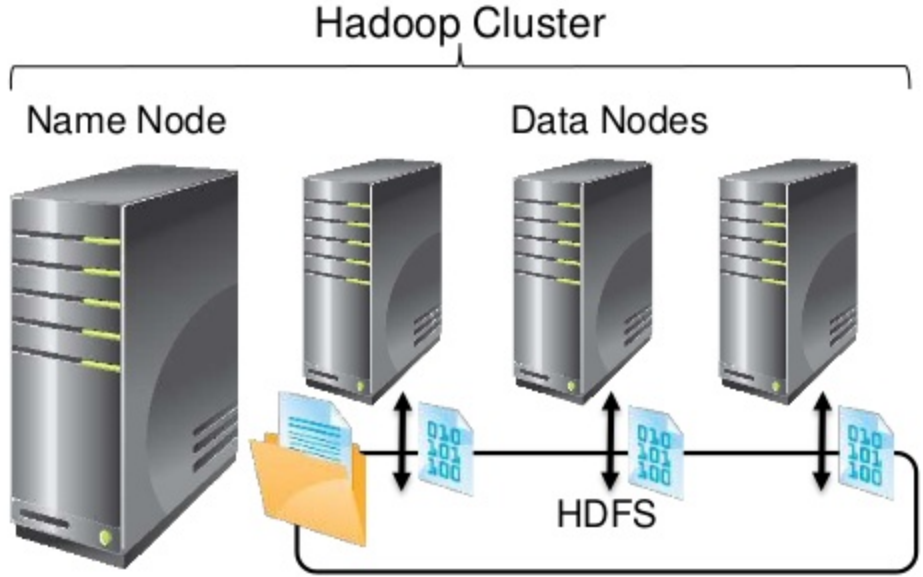
## Hadoop almacena archivos

Los archivos pueden ser semi estructurados o no estructurados



# Apache Hadoop

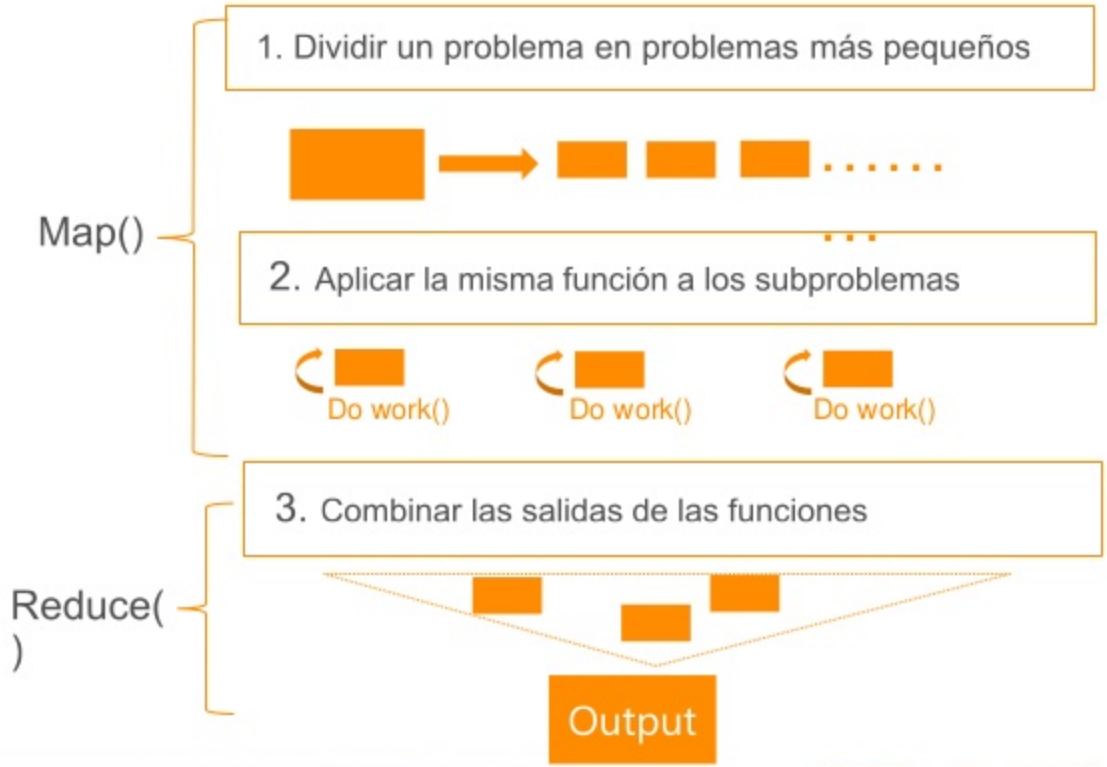
- Hadoop Cluster
  - Múltiples servidores con Sistema de archivos compartido (HDFS)
  - Name Node que atiende las peticiones de los clientes
  - Múltiples nodos de datos que utilizan Map Reduce



# Map Reduce

- Estrategia divide y vencerás:

- Map() – divider en problemas más pequeños
- Reduce() – combinar los resultados



# Big Data sobre Azure

- Infraestructura como un Servicio(IaaS)
  - ❑ Hadoop en una VM
    - ✓ Hortonworks
    - ✓ Cloudera
    - ✓ MapR

- Plataforma como un Servicio(PaaS)
  - ❑ Azure HDInsight(Cluster as a Service)
  - ❑ Azure Data Lake Store and Analytics

# HDInsight


Batch
Map Reduce
Script
Pig
SQL
Hive
NoSQL
HBase
Streaming
Storm
In-Memory
Spark
Predictive
R Server
Core Engine

# Que es un Data Lake?

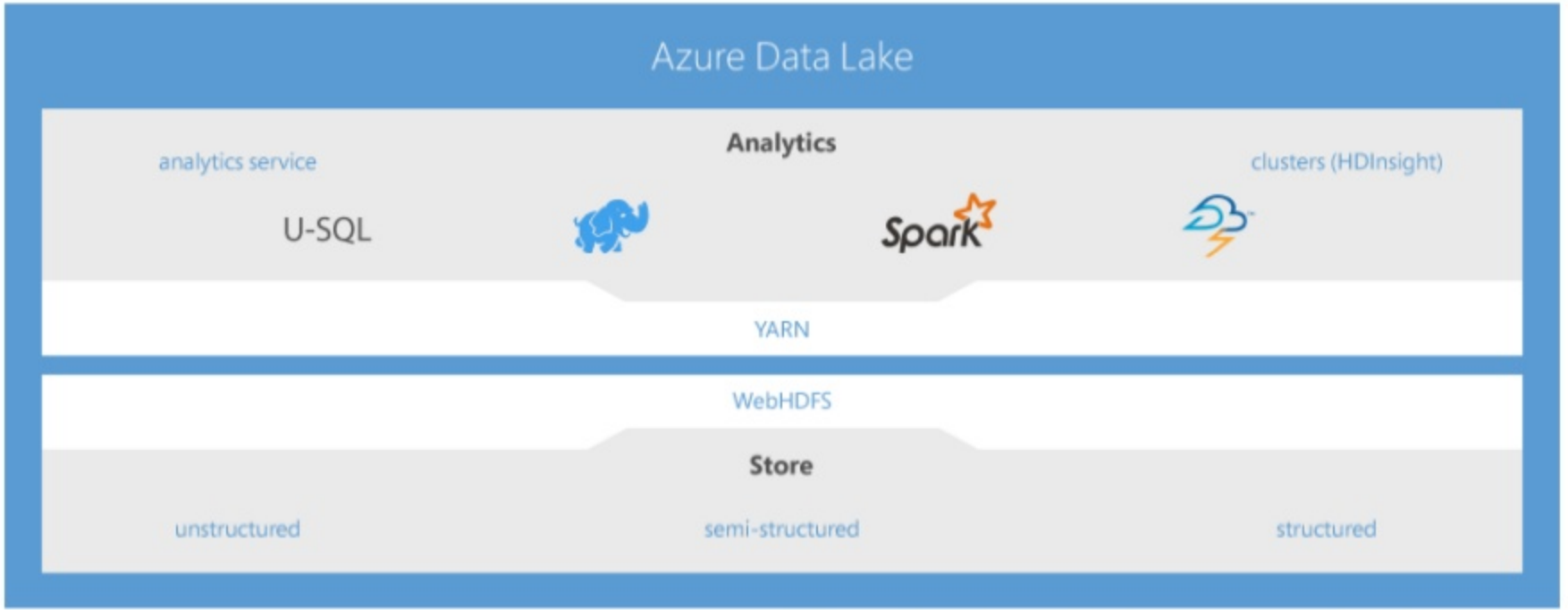
*"Un simple almacenamiento de toda la data...desde la data cruda(que implica una copia exacta del origen de datos)*

*a los datos transformados que son usados de varias formas incluyendo reportes, visualizaciones, analítica y maquinas de aprendizaje."*

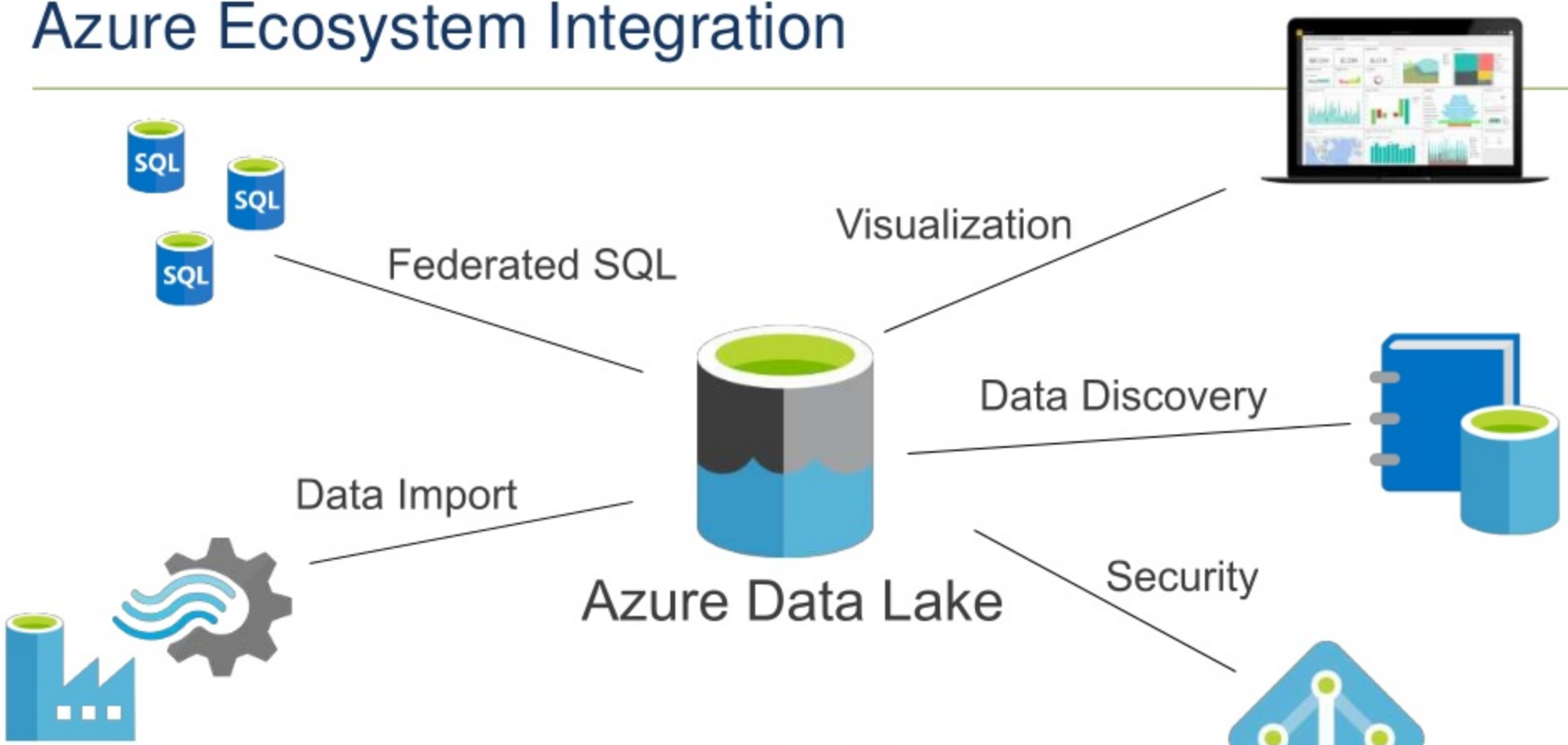
# Azure Data Lake

- Integrando, plataforma de Big Data Storage + Analytics
- Diseñado de las experiencias del mundo real
  - ✓ Office 365, Skype, Bing, etc.
- Aprovechar tecnologías y habilidades existentes
- Beneficios de un Servicio Local Azure
  - ✓ Elasticidad, aprovisionando dinámicamente los recursos que necesitamos
  - ✓Capacidad de almacenamiento infinito
  - ✓Enfocado en extracción significante de la data, no en la infraestructura

# Built on Open-Source



# Azure Ecosystem Integration



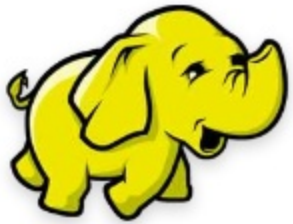
# Azure Data Lake Store

- HDFS como servicio
- Almacenamiento durable
- Una variedad de escenarios
  - Alta Capacidad
  - Alta Frecuencia
  - Alto Rendimiento
- Data se almacena en su formato nativo
  - Formatos de almacenamiento estructurado, semiestructurado y no estructurado

# Azure HDInsight


- Administrado, nube escalable de Hadoop como un Servicio
- Complemento complete de las tecnologías de Apache.

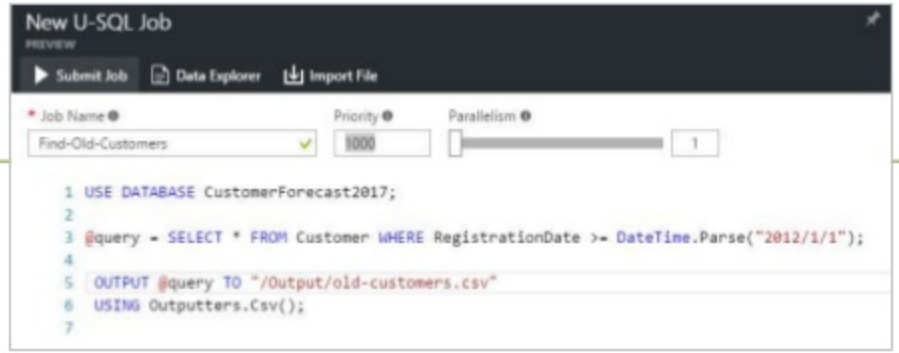
  Spark, Storm, HBase, etc.
- Se centran en consultas y datos, no infraestructura
- Pagas por solo lo que necesitas usar
- Aprovechar las herramientas existentes
  - Hive, Pig, Sqoop, R, etc.



# Azure Data Lake Analytics

- Complemento al ecosistema HDInsight y Hadoop.
- Lo escalas dinámicamente para coincidir con complejidad de tamaño y consulta de datos
- Construido en Apache YARN
- Unidad de interacción es un trabajo de análisis.
- U-SQL: Lenguaje de consulta arraigada entre SQL y C#

# U-SQL




- Basado en SQL y C#
  - Tipos y expresiones C#
  - Tablas, vistas, funciones de Windows.
  - Funciones definidas por el usuario/operadores/agregaciones en C.
- Trabajo típico
  1. Leer la data de archivos/tabla/ origenes federados
  2. Transforma las filas en un pipeline.
  3. Filas de salida a tablas o filas.

# Read the input, write it directly to output (just a simple copy)

```
@orders =
      EXTRACT
          OrderId  int,
          Customer string,
          Date     DateTime,
          Amount   float
      FROM "/input/orders.txt"
      USING Extractors.Tsv();

OUTPUT @orders
    TO "/output/orders_copy.txt"
    USING Outputters.Tsv();
```

Rowset

Apply Schema on read

From a file in an ADL Store

Easy delimited text handling

Write out

# Transforming Rowsets

```
@customers =
    SELECT Customer.ToUpper() AS Customer
    FROM @orders
    WHERE Customer.Contains("Contoso");
```

C# Expression

C# Expression

Use WHERE for filtering

PASS SQL saturday

# Refining Rowsets

```
@orders =
    SELECT *
    FROM @orders
    WHERE Customer.Contains("Contoso");
```

# Use Your own C# methods

```
@rows =
    SELECT
        OrdersDB.Helpers.Normalize(Customer) AS Customer,
        Amount AS Amount
    FROM @orders;
```

Use your own helper functions

# Grouping & Aggregation

```
@rows =
    SELECT
        Customer,
        SUM(Amount) AS TotalAmount
    FROM @orders
    GROUP BY Customer;
```

Many other aggregations are possible. You can define your own aggregator with C#!

# Grouping & Aggregation (2)

```
@rows =
    SELECT
        Customer,
        SUM(Amount) AS TotalAmount
    FROM @orders
    GROUP BY Customer
    HAVING TotalAmount > 1000000;
```

HAVING filters the output of a GROUP BY

# DECLARE values for later use

```
DECLARE @text1 string = "Hello World";
DECLARE @text2 string = @"Hello World";
DECLARE @text3 char = 'a';
DECLARE @text4 string = "BEGIN" + @text1 + "END";
DECLARE @text5 string = string.Format("BEGIN{0}END", @text1);

DECLARE @numeric1 sbyte = 0;
DECLARE @numeric2 short = 1;
DECLARE @numeric3 int = 2;
DECLARE @numeric4 long = 3L;
DECLARE @numeric5 float = 4.0f;
DECLARE @numeric6 double = 5.0;

DECLARE @d1 DateTime = System.DateTime.Parse("1979/03/31");
DECLARE @d2 DateTime = DateTime.Now;

DECLARE @misc1 bool = true;
DECLARE @misc2 Guid = System.Guid.Parse("BEF7A4E8-F583-4804-9711-7E608215EBA6");
DECLARE @misc4 byte [] = new byte[] { 0, 1, 2, 3, 4};
```

text

numeric

Date/time

Other

# Creating Constant Rowsets in Script

```
@departments  =
    SELECT * FROM
        (VALUES
            (31,    "Sales"),
            (33,    "Engineering"),
            (34,    "Clerical"),
            (35,    "Marketing")
        ) AS
              D( DepID, DepName );
```

# Sorting a rowset

```
@customers
    SELECT *
    FROM @customers
    ORDER BY Amount ASC
    FETCH FIRST 3 ROWS;
```

SELECT with ORDER BY requires a FETCH FIRST!

PASS SQL saturday

# Sorting on OIUTPUT

```
OUTPUT @customers
    TO @"/output.tsv"
    ORDER BY Amount ASC
    USING Outputters.Tsv();
```

# DEMO

PASS SQL saturday

#608 | BOGOTÁ 2017

#sqlsatBogota


Reportes 360 en mi Organización



PASS SQL saturday

PASS SQL saturday

#608 | BOGOTÁ 2017

#sqlsatBogota


Preguntas y Respuestas



PASS SQL saturday

PASS SQL saturday

#608 | BOGOTÁ 2017

**“En Dios confiamos, todos los demás traigan datos”**

Gracias !!!