SolidQ

The H Hour: Hadoop
The awakening of the BigData

Antonio Soto

SolidQ COO

asoto@solidq.com

@antoniosql

Tendencias de la Industria

EXPLOSIÓN DE DATOS	CONSUMERIZACIÓN DE TI	NUBE PÚBLICA Y PRIIVADA

Los datos digitales crecerán **44x** próxima década

En 2015, servicios de nube pública tendrán **46%** de crecimiento neto en gasto de ti

IDC Digital Universe Study, sponsored by EMC, May 2010.
Worldwide and Regional Public IT Cloud Services 2011 – 2015 Forecast, IDC , June 2011—Doc #228485

El nuevo rol del operador

El operador de ayer	El operador de hoy
Sigue el proceso basado en procedimientos predefinidos	Toma decisiones objetivas basadas en datos en tiempo real
Trabajar dentro de una función lineal y funcional	Trabaja en una organización interfuncional
Mantener el cumplimiento de las normas de ajuste	Contribuir a la conducción de cambios de procesos
Tomar decisiones independientes basadas en formación	Aprovechar el conocimiento institucionalizado

Agenda

- ¿Qué es Big Data?
- Entonces… Hadoop, ¿Qué es?
  - Ventajas
  - Componentes
- Apache Hadoop y Microsoft BI
  - HDInsight
  - Windows Azure HDInsight
- Casos de Uso

¿Qué es Big Data?

Big data Consists of datasets that grow so large that they become awkward to work with using on-hand DB Management tools.

Wikipedia

Big data is when the size of the data itself becomes part of the problem

Mike Lukides, O'Reilly Radar

It's not just your "Big Data" problems, it's all about your BIG "data" Problems.

Alexander Stojanovic, Hadoop Manager on Win Azure

Las 4 V's

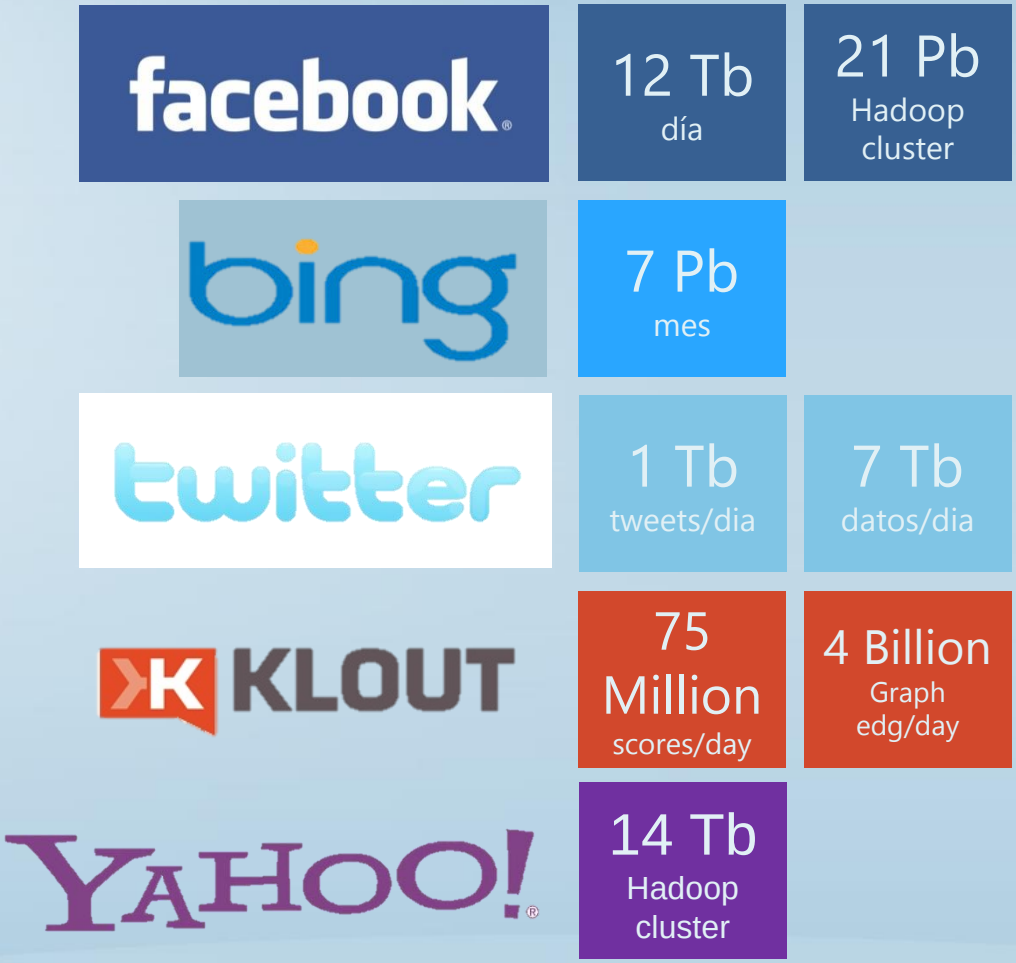
Volumen

Velocidad

Variedad

Variabilidad

Ejemplos de Big Data



Entonces…¿cómo obtengo insights?

Datos estructurados

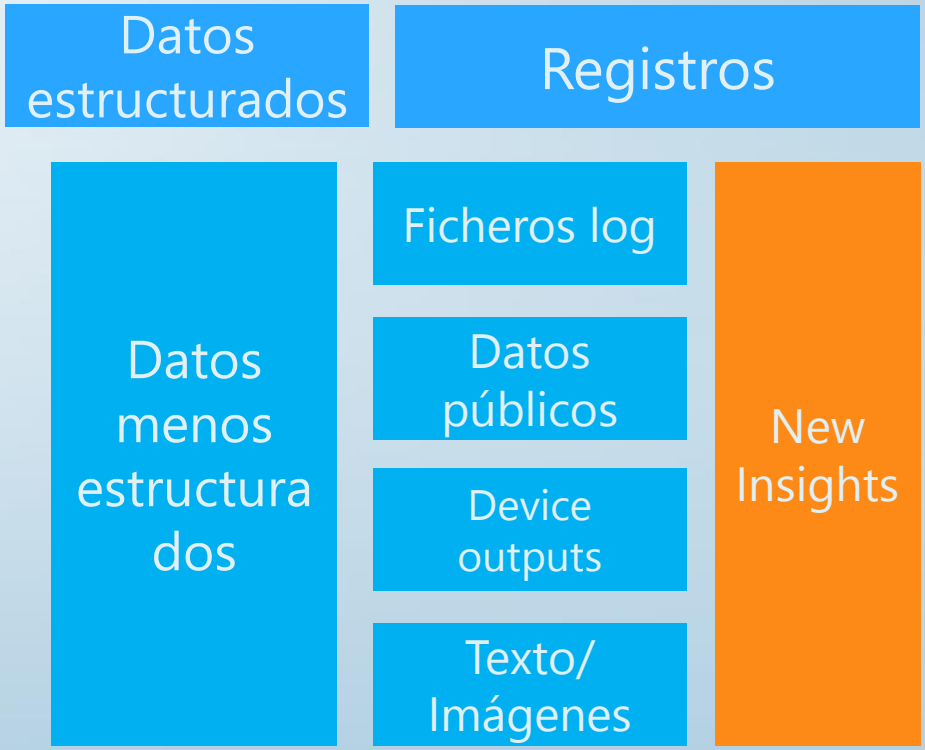
Registros

- Datos estructurados
  - Bases de Datos relacionales
  - Bases de Datos analíticas

Entonces…¿cómo obtengo insights?

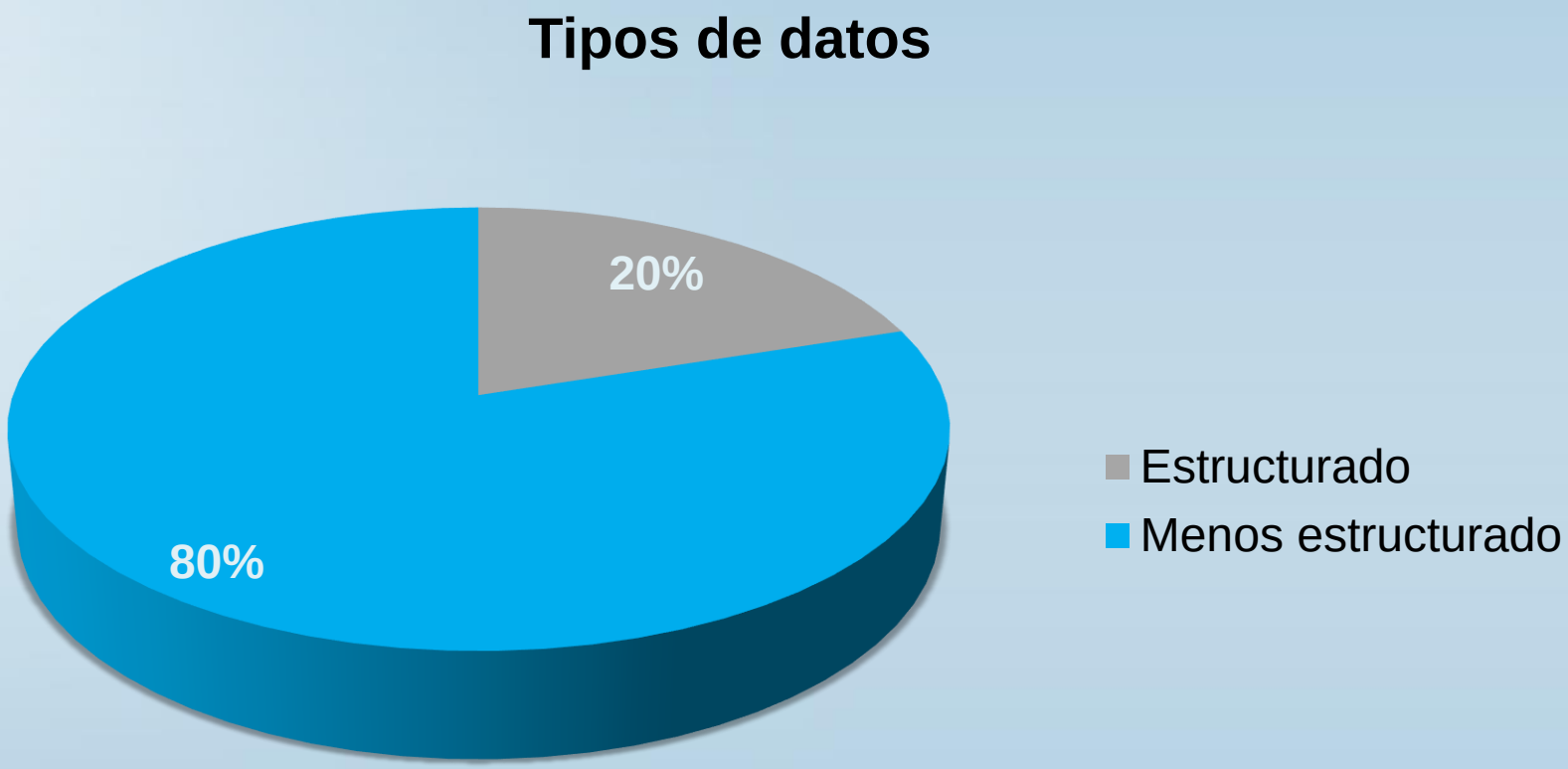
- Datos estructurados
  - Bases de Datos relacionales
  - Bases de Datos analíticas

Entonces…¿cómo obtengo insights?



- Datos estructurados
  - Bases de Datos relacionales
  - Bases de Datos analíticas
- Datos menos estructurados
  - Intentar un ETL para transformarlo en relacional
    - Tiempo de desarrollo elevado
    - Son datos susceptibles a cambios de estructura
  - Archivados y Borrados
  - Acceso caro

Entonces…¿cómo obtengo insights?



Insights de datos no estructurados

DEMO

¿Qué es Hadoop?

- Open Source
- Plataforma de almacenamiento de datos y análisis para Big Data
- Optimizado para manejar
  - Datos masivos a través de paralelismo
  - Variedad de datos (Estructurados, No-estructurados, Menos estructurados)
  - Uso de hardware económico
- No para OLTP / OLAP

¿Qué es Hadoop?: Ventajas

Escalable

Escala linealmente en capacidad de almacenamiento y computación

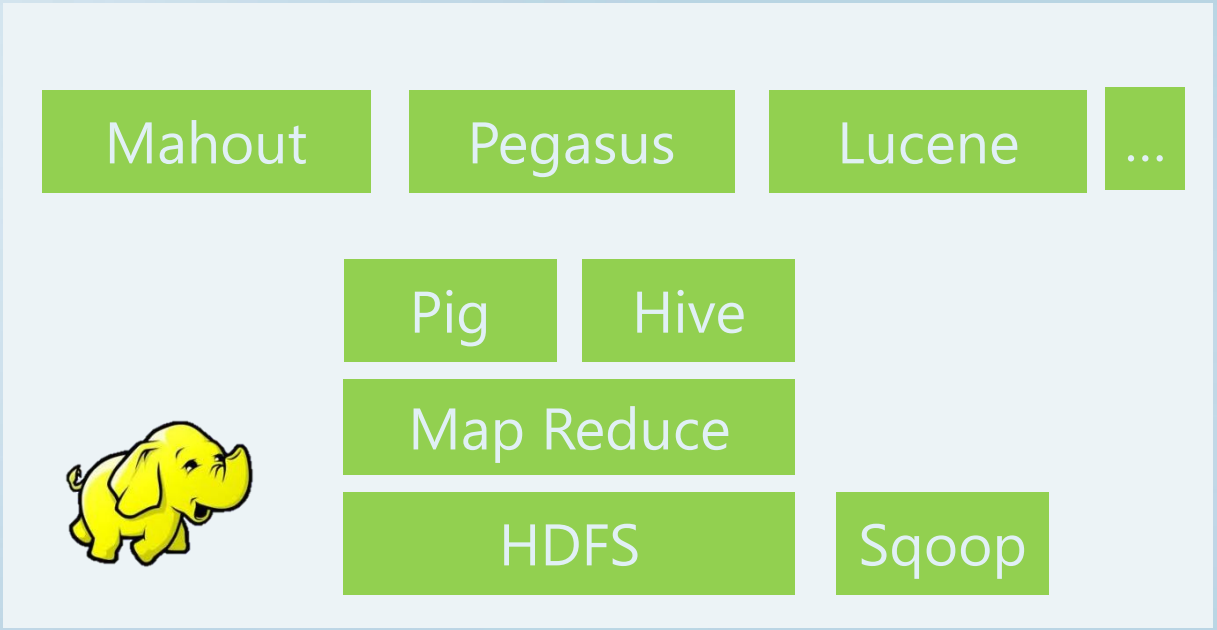
Tolerante a Fallos

Proporcionado por el Sistema de ficheros distribuido y el framework de lectura

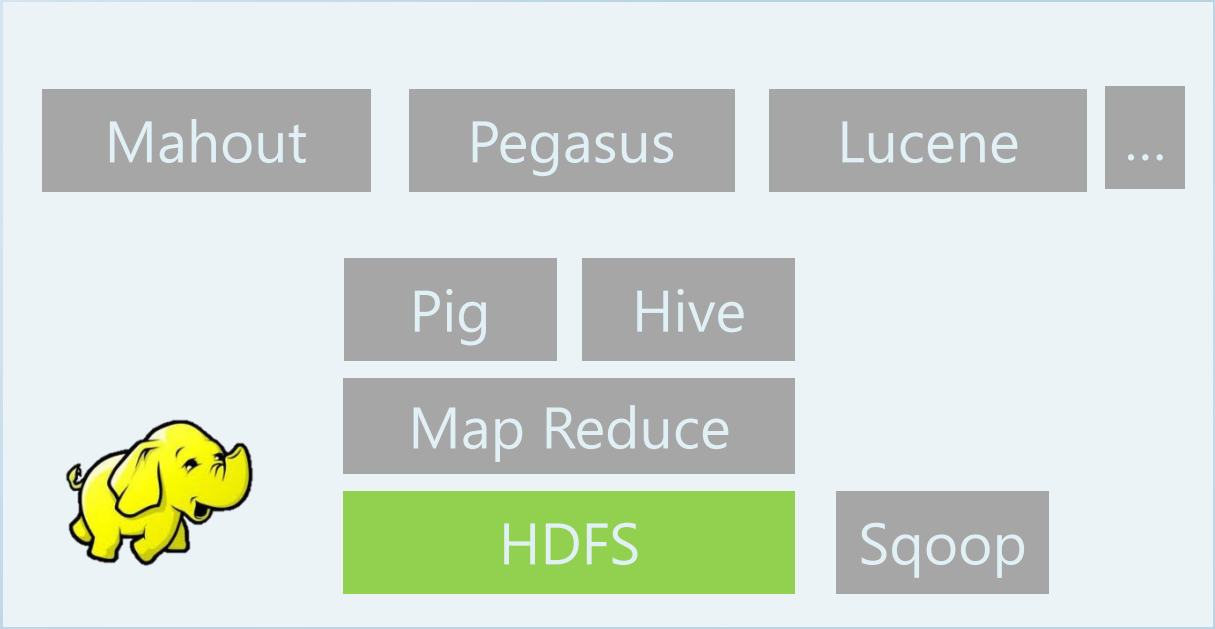
Procesamiento distribuido

Sigue la estrategia de divide y vencerás

¿Qué es Hadoop?: Componentes



¿Qué es Hadoop?: Componentes



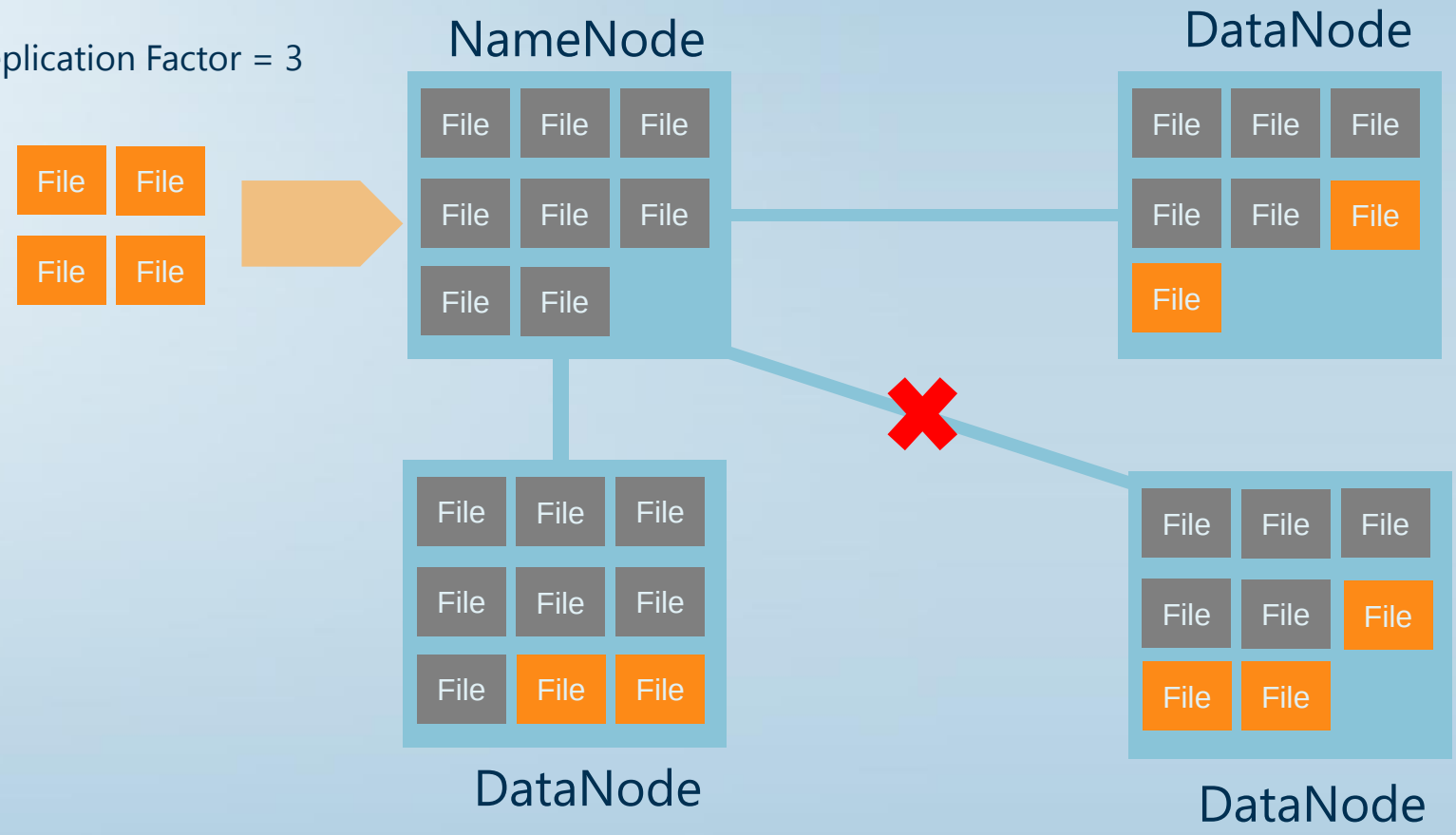
Hadoop Distributed File System (HDFS)

- Sistema de ficheros distribuido diseñado para grandes conjuntos de datos
- Fiable y con buen rendimiento
  - Alto rendimiento de acceso: Latencia de disco
  - Alto ancho de banda Almacenamiento Clustered auto-reparable
- Divide los datos entre los nodos en un Cluster
  - NameNode: Mantiene el mapeo de bloques de ficheros a nodos esclavos
  - DataNode: Almacena y sirve bloques de datos

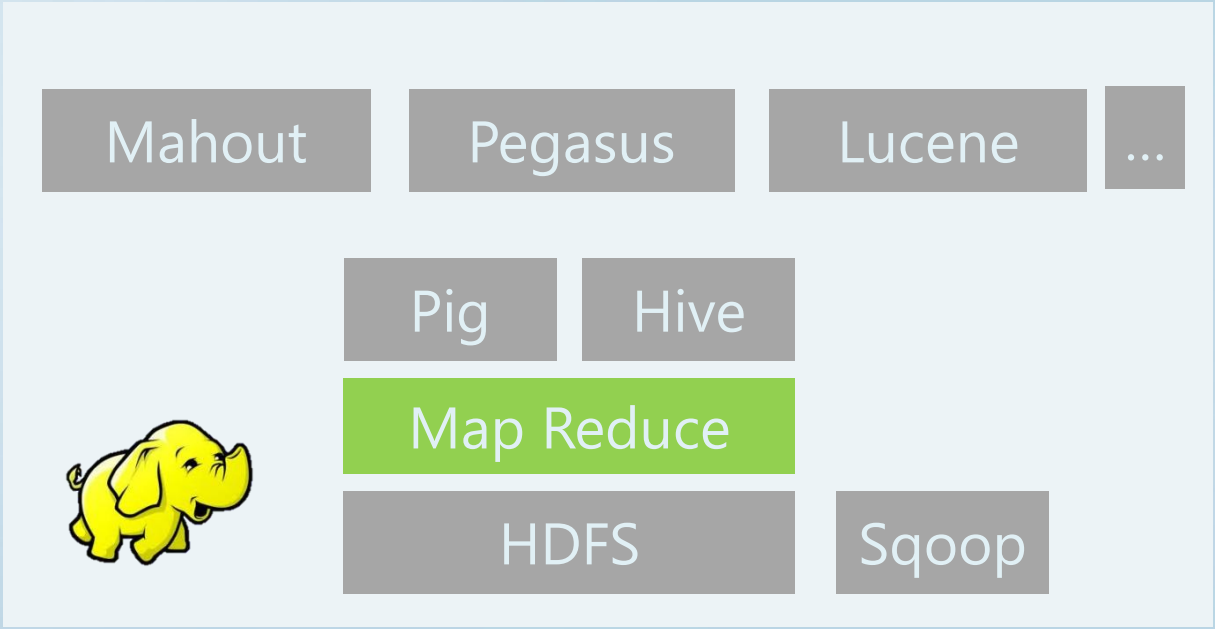
Hadoop Distributed File System (HDFS)

Block Size = 64 Mb

Replication Factor = 3



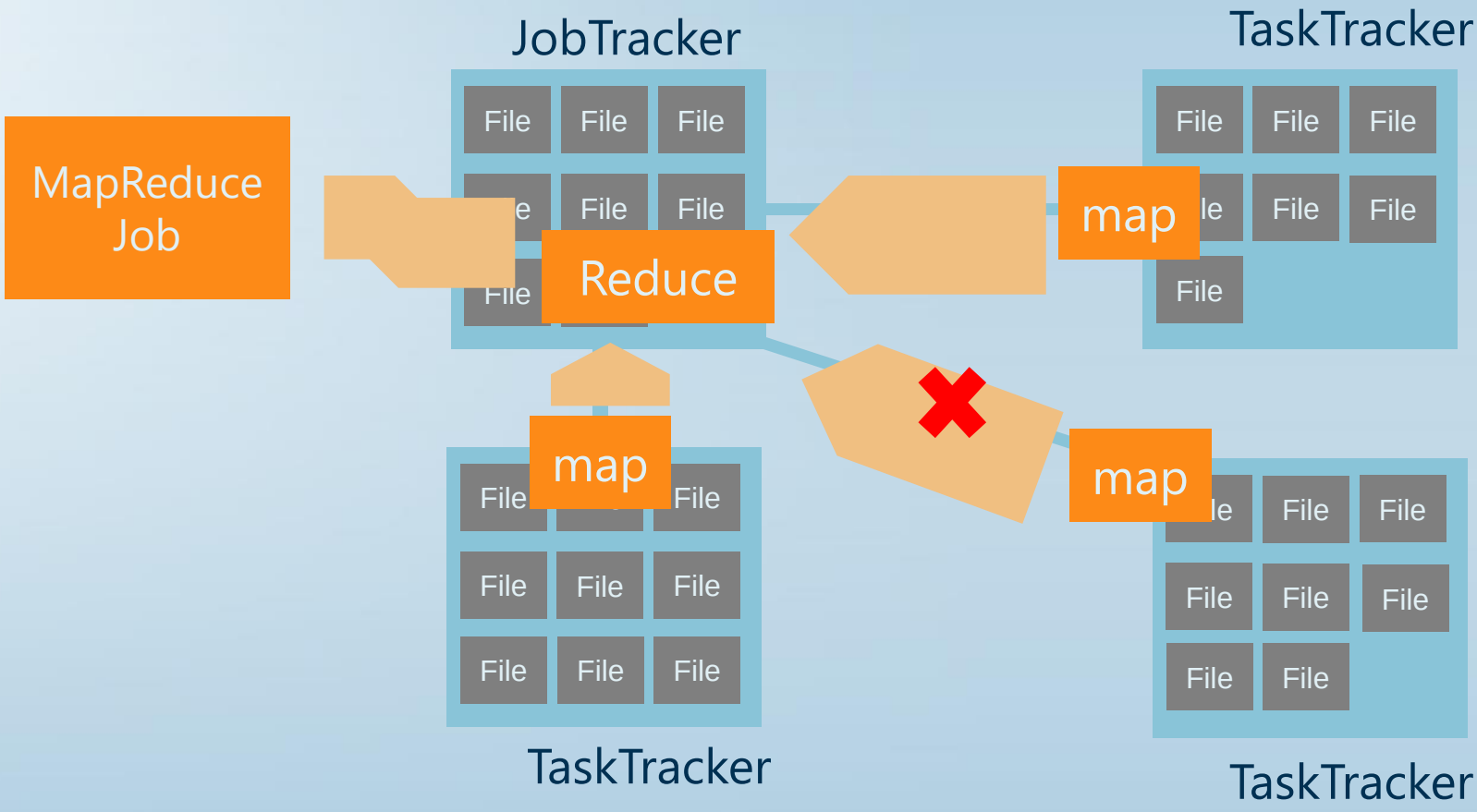
¿Qué es Hadoop?: Componentes



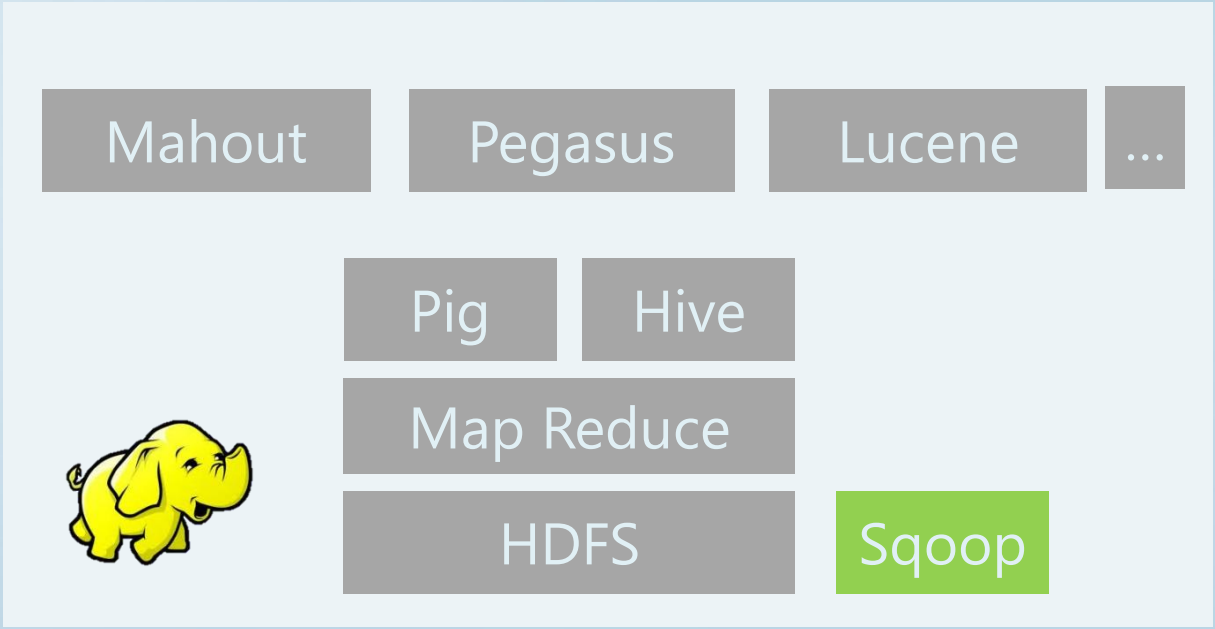
Map Reduce Framework

- Motor de planificación parar Procesamiento de carga distribuido
  - Pares Clave-Valor
  - Función Map
  - Función Reduce
- Lenguajes de Script : Java, python, Javascript…
- Saca provecho de la distribución de datos de HDFS
  - JobTracker: Planifica los trabajos entre los TaskTrackers
  - TaskTracker: unidades de trabajo

Map Reduce Framework



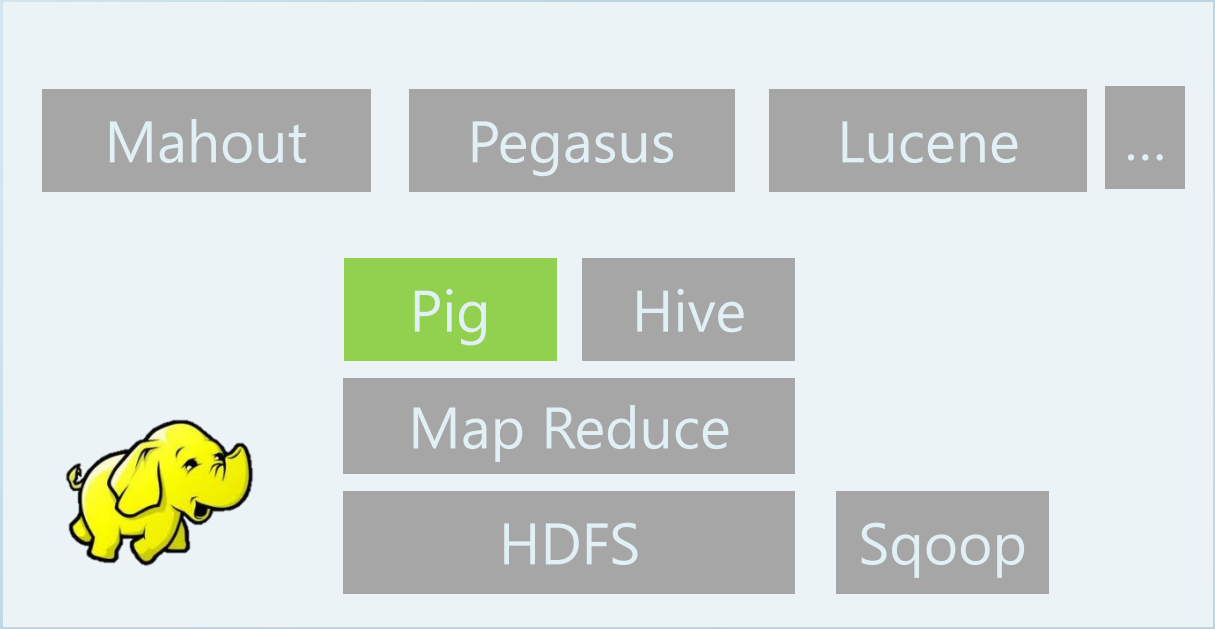
¿Qué es Hadoop?: Componentes



Sqoop

- Tecnología que sirve de interfaz entre HDFS y los Sistemas de información empresarial
- Orígenes de datos relacionales integrados
  - MySQL, Oracle, SQL Server …
- Importación / Exportación (Bidireccional)

¿Qué es Hadoop?: Componentes



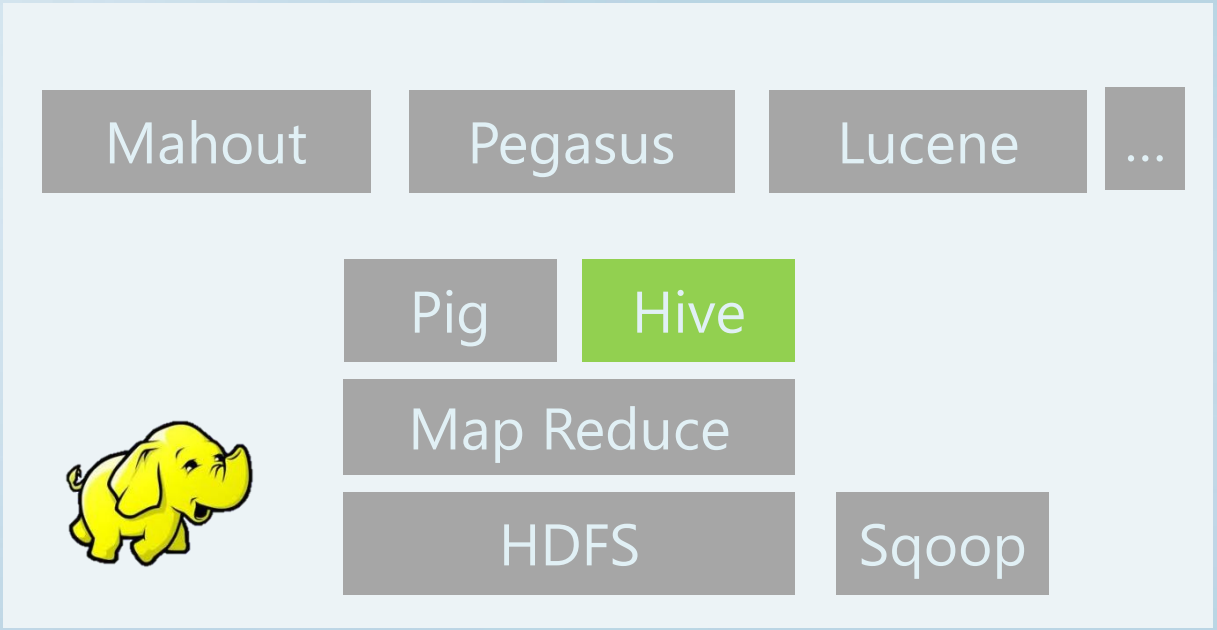
Pig

- Lenguaje de flujo de datos de alto nivel y framework de ejecución
- Lenguaje de consulta: PigLatin
  - Posibilidad de join de tablas

```
log  = LOAD 'excite-small.log' AS (user, time, query);
grpd = GROUP log BY user;
cntd = FOREACH grpd GENERATE group, COUNT(log);
DUMP cntd;
```

- Por detrás ejecuta trabajos MapReduce

¿Qué es Hadoop?: Componentes



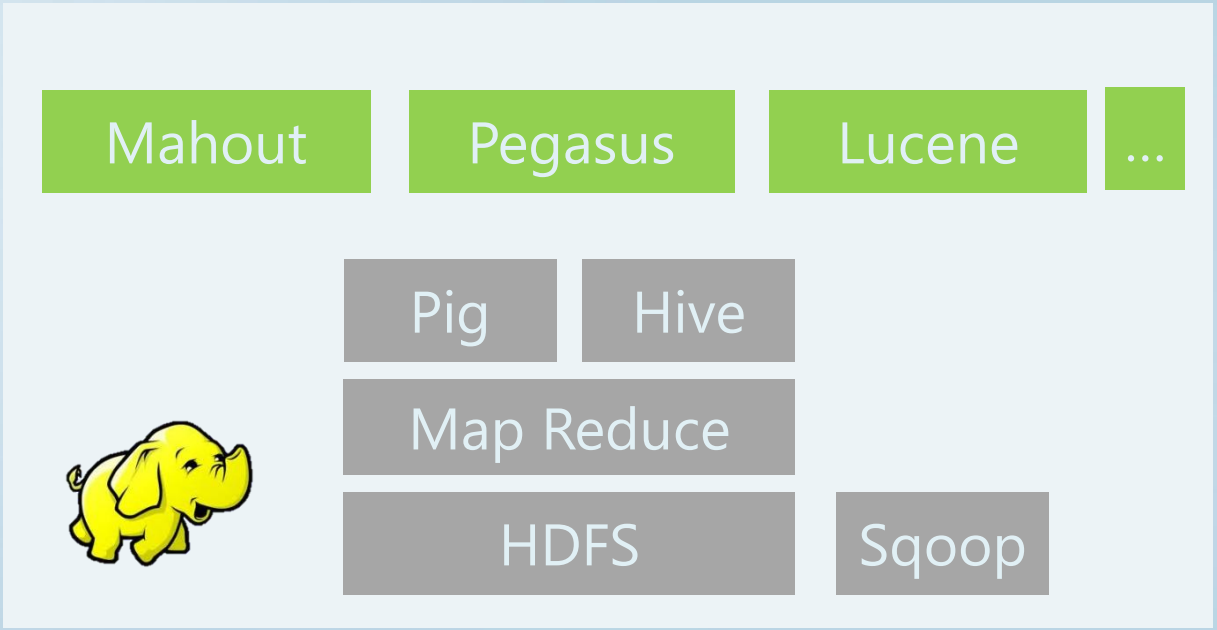
Hive

- Infraestructura Data Warehouse desde Hadoop
- Proporciona
  - Sumarización de Datos
  - Consultas Ad-hoc
- Lenguaje consulta estilo SQL: HiveQL

```
select regexp_replace(split(csuristem, "/")[1], "MainFeed.aspx", "Home"),
count(*)
from weblog_sample
group by regexp_replace(split(csuristem, "/")[1], "MainFeed.aspx", "Home")
```

- Por detrás ejecuta trabajos MapReduce

¿Qué es Hadoop?: Componentes



Otros componentes: Hadoop Ecosystem

Mahout

- Minería de Datos y Machine Learning

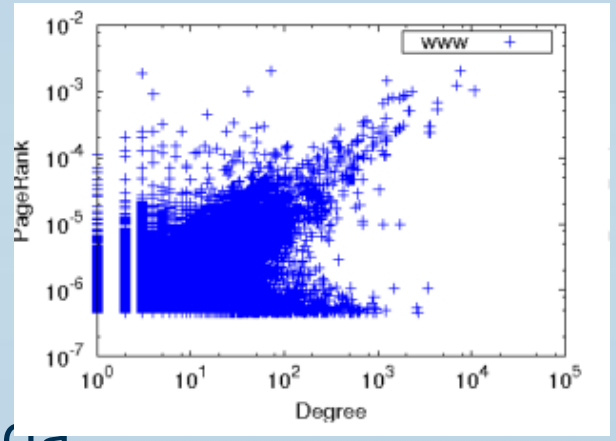
Pegasus

- Page Rank yGraph Mining
- Social Network Analysis

Lucene

- Technología de indexación y búsqueda

Algunos otros: Avro, Hbase, Flume, Oozie…

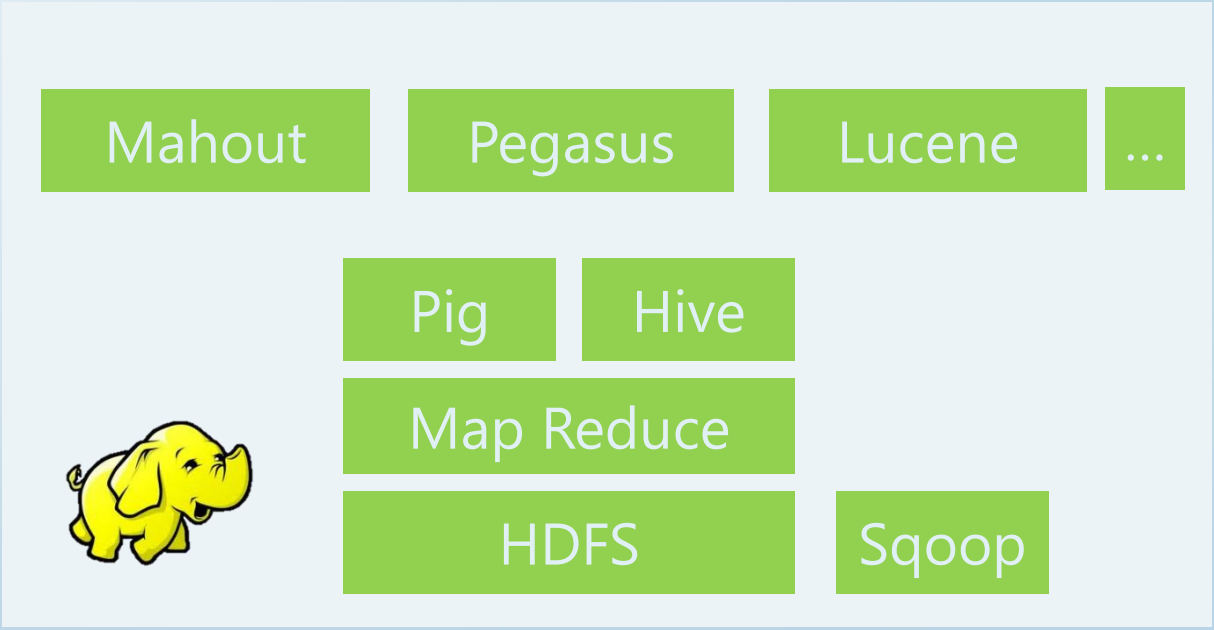


MICROSOFT ON THE HADOOP

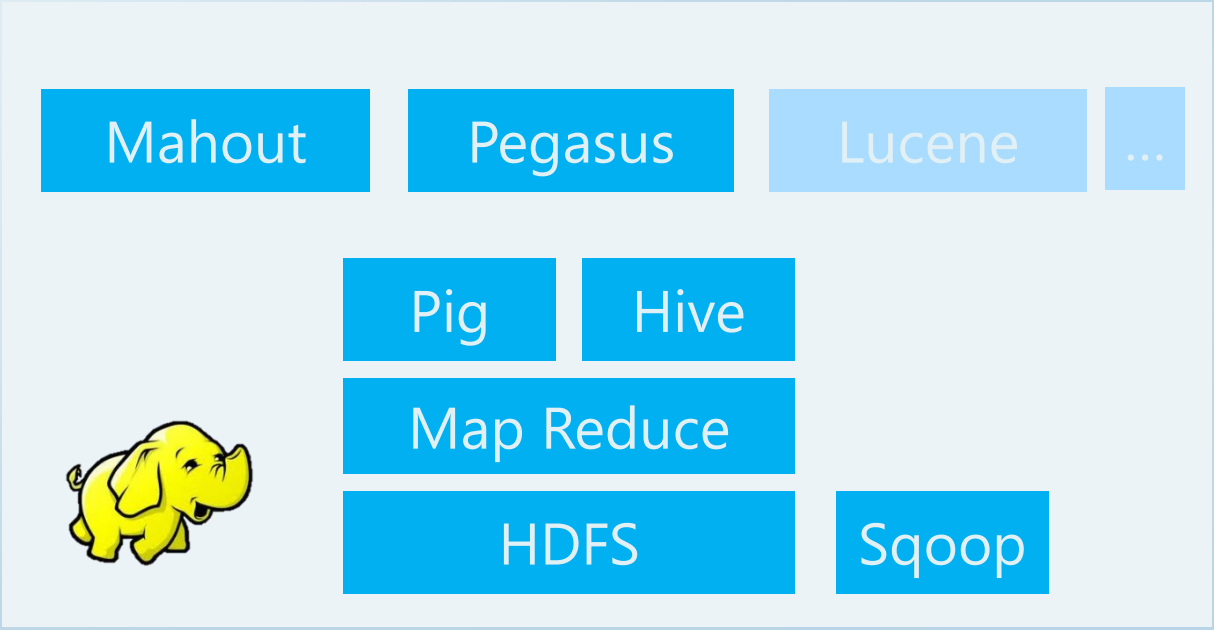
HDInsight

- Project Isotope
- Proporciona Apache Hadoop en
  - Windows Server
  - Windows Azure
- Active Directory & System Center

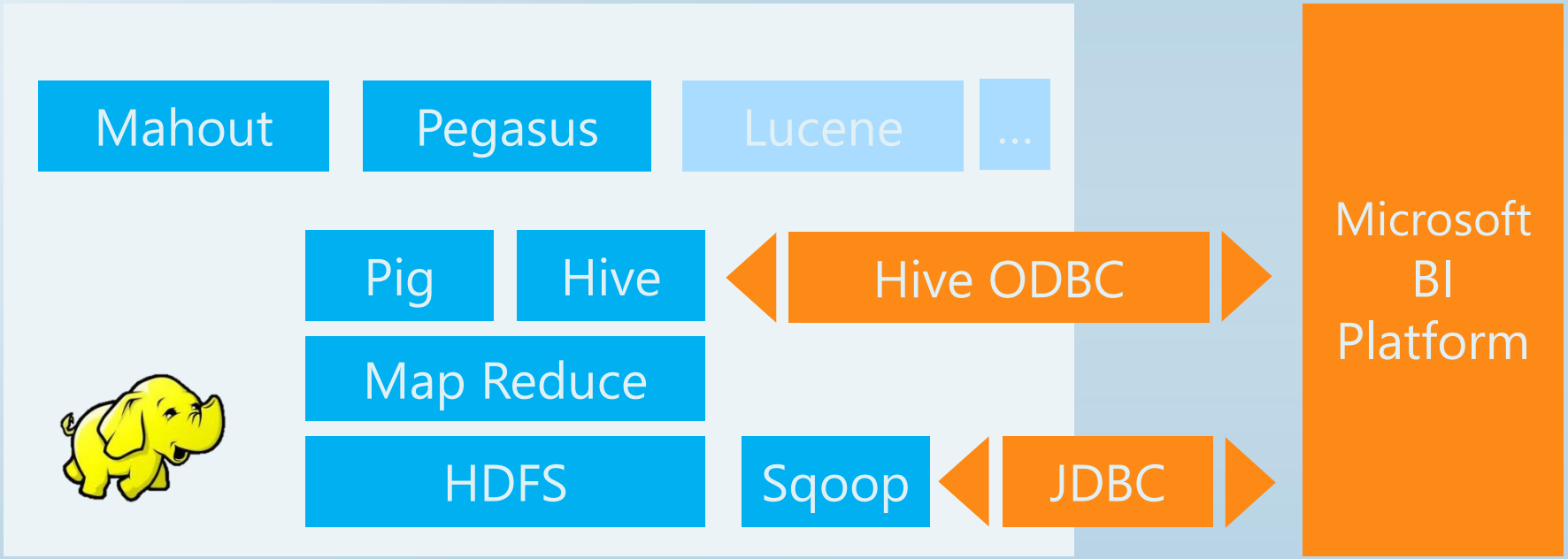
Hadoop: Componentes Originales



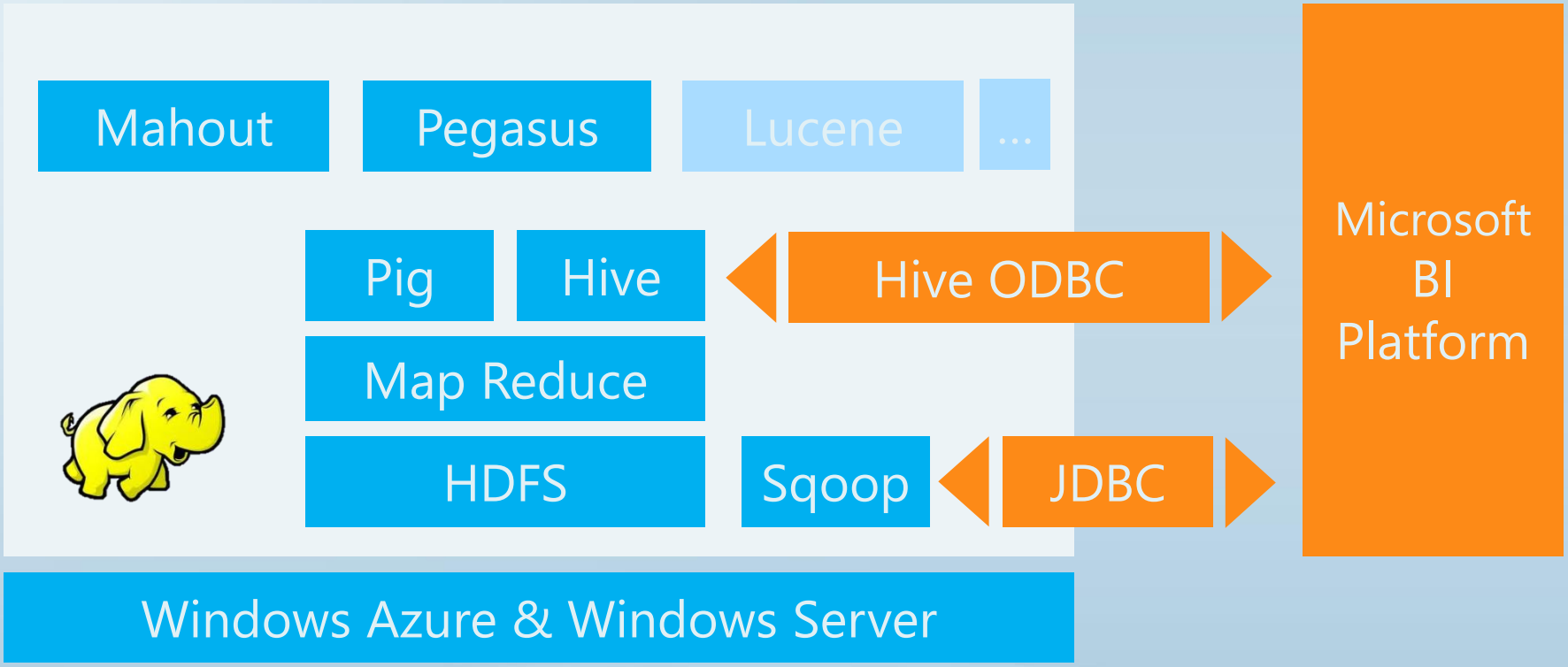
HDInsight



HDinsight


Mahout
Pegasus
Lucene
...
Pig
Hive
Hive ODBC
Map Reduce
HDFS
Sqoop
JDBC
Microsoft BI Platform

HDInsight


Mahout
Pegasus
Lucene
…
Pig
Hive
Hive ODBC
Map Reduce
HDFS
Sqoop
JDBC
Microsoft BI Platform
Windows Azure & Windows Server

Windows Azure HDInsight

DEMO

Características HDInsight

HDFS

- Basado en Windows
- Compatibilidad con Directorio Activo
- Almacenamiento compatible:
  - HDFS
  - Azure Blob Storage
  - Amazon S3

MapReduce Framework

- Compatibilidad JavaScript
- Hadoop Streaming con compatibilidad F# y C#

Características HDInsight

Hive

- Consolta Interactiva
- Complemento Hive para Excel 2010
- Hive ODBC Driver
- Potentes funciones regex

Pig

- Consola Interactiva

Sqoop

- Driver JDBC para SQL Server y SQL Server PDW

Trabajando con HDInsight

DEMO

Casos de Uso

- Analítica de Eventos
- Analítica de clics a gran escala
- Optimizaciones de precio
- Gestión de riesgo financiero
- Análisis de sentimiento
- Minería de datos a gran escala

Recapitulando

- HDInsight nos permite almacenar, procesar y analizar datos menos estructurados
- Los proyectos de Apache Hadoop Ecosystem agregan características extra
- Complementa y enriquece el Análisis de Negocio
- Encaja perfectamente con la Experiencia Cloud

PREGUNTAS

Gracias!

Antonio Soto

asoto@solidq.com

@antoniosql