

# Introducción Apache Hadoop

**Meetup Santiago, Chile**

Junio 26, 2018



# Presentador

**Italo Cocio**

*Technical Support Engineer*

*Hortonworks Support Team Chile.*

# Agenda

- ◆ Qué es Big Data?
- ◆ Problemática de los datos
- ◆ Qué es Apache Hadoop?

# Qué es Big Data?

# Qué es Big Data?

- Big Data se refiere al conjuntos de datos cuyo tamaño, complejidad y velocidad de crecimiento dificultan su captura, gestión, procesamiento o análisis mediante tecnologías convencionales, tales como bases de datos relacionales, dentro del tiempo necesario para que sean útiles.



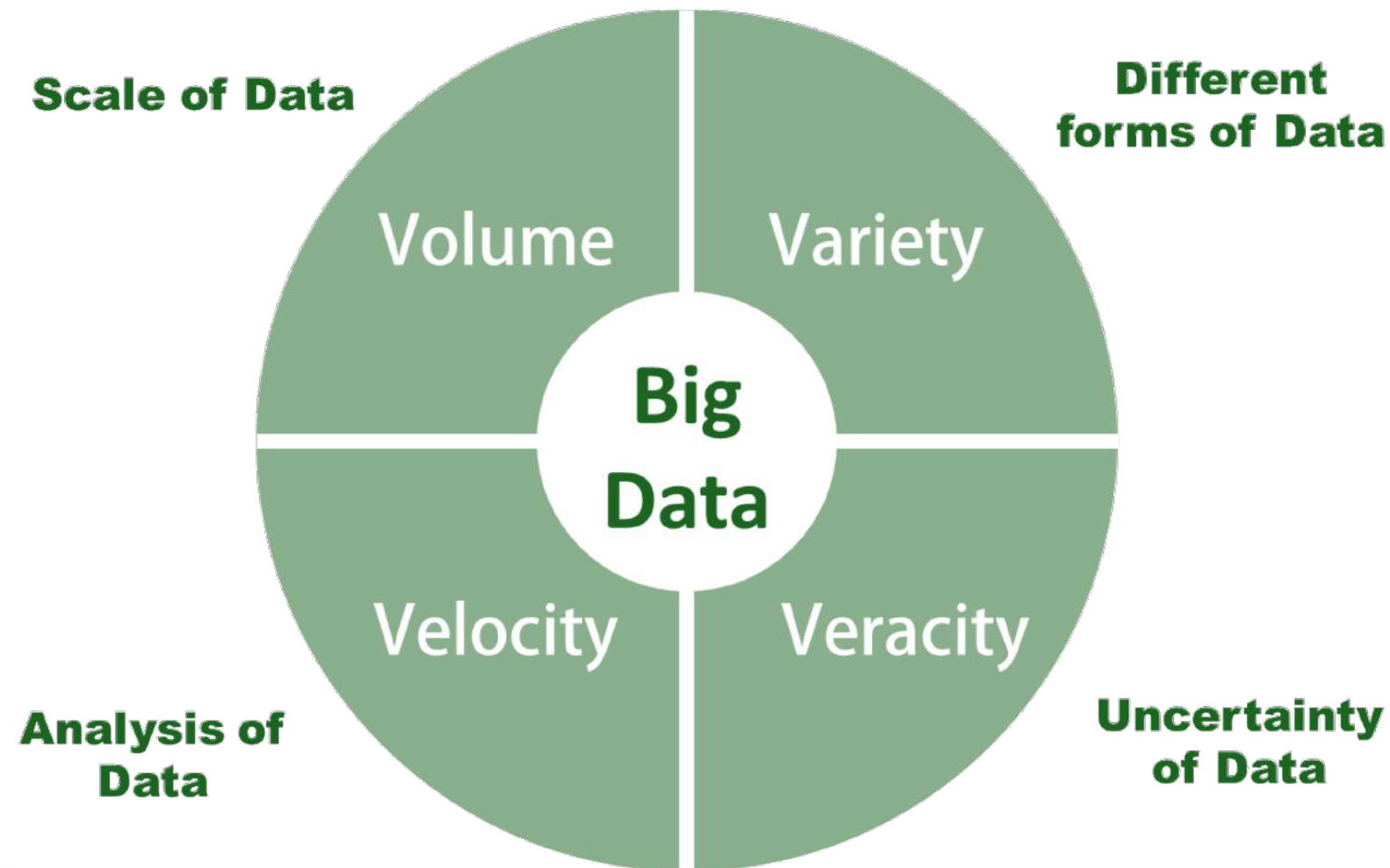
# Qué es Big Data?

- ◆ 294 mil millones de emails enviados cada día.
- ◆ Sobre 1.000 millones de búsquedas en Google cada día.
- ◆ Más de 30 Petabytes de datos generados, almacenados, accedidos y analizados por Facebook.
- ◆ Más de 230 millones de tweets generados cada día.

# Qué es Big Data?

- ◆ Cientos de miles de millones de sensores monitoreando, rastreando y comunicándose entre sí, para poblar la IoT con datos en tiempo real.
- ◆ International Data Corporation (IDC), estima que para el año 2020, las transacciones comerciales via internet (incluyendo B2B y B2C) alcanzará los 450 mil millones por día.
- ◆ Para el año 2020, se estima que la información generada por cada ser humano alcanzará aprox. 1.7 megabytes por segundo.

# Características de los datos





# Problemática de los datos

# Problemática de los datos

- ◆ Las arquitecturas existentes hacen innaccesible , incompleta, irrelevante y costosa la disponibilidad de la información.
- ◆ Debido a la generación acelerada de información:
  - Costos de almacenamiento
  - Reformato
  - Consulta

crecen más rápido que el valor que los datos mismos pueden proveer.

# Qué es Apache Hadoop?

# Qué es Apache Hadoop?

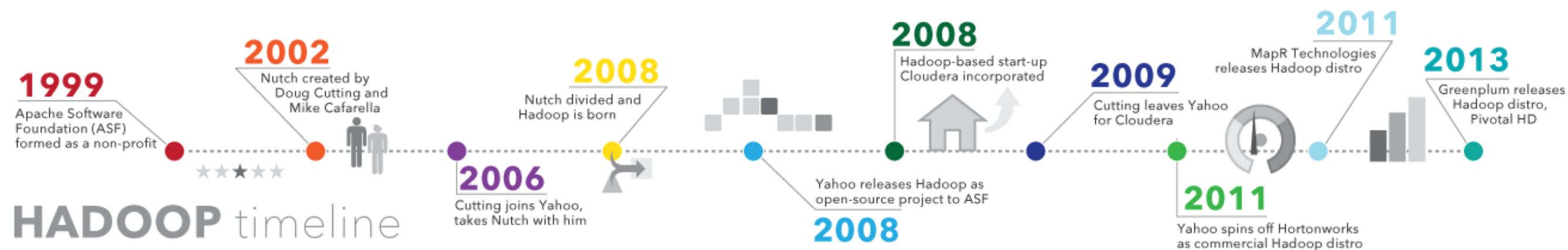
- ◆ Es un plataforma de software open source para almacenamiento y procesamiento distribuido de grandes cantidades de datos en cluster contruídos desde lo que llamamos "commodity hardware".
- ◆ Solución para la problemática.
- ◆ Permite lidiar con complejidades, como el gran volumen y variedad de datos.
- ◆ Es un conjunto de proyectos open source.
- ◆ Uso de commodity hardware para almacenamiento.

# Qué es Apache Hadoop?

- ◆ Es un plataforma de software open source para almacenamiento y procesamiento distribuido de grandes cantidades de datos en cluster contruídos desde lo que llamamos "commodity hardware".
- ◆ Solución para la problemática.
- ◆ Permite lidiar con complejidades, como el gran volumen y variedad de datos.
- ◆ Es un conjunto de proyectos open source.
- ◆ Resiliencia
- ◆ Uso de commodity hardware para almacenamiento.

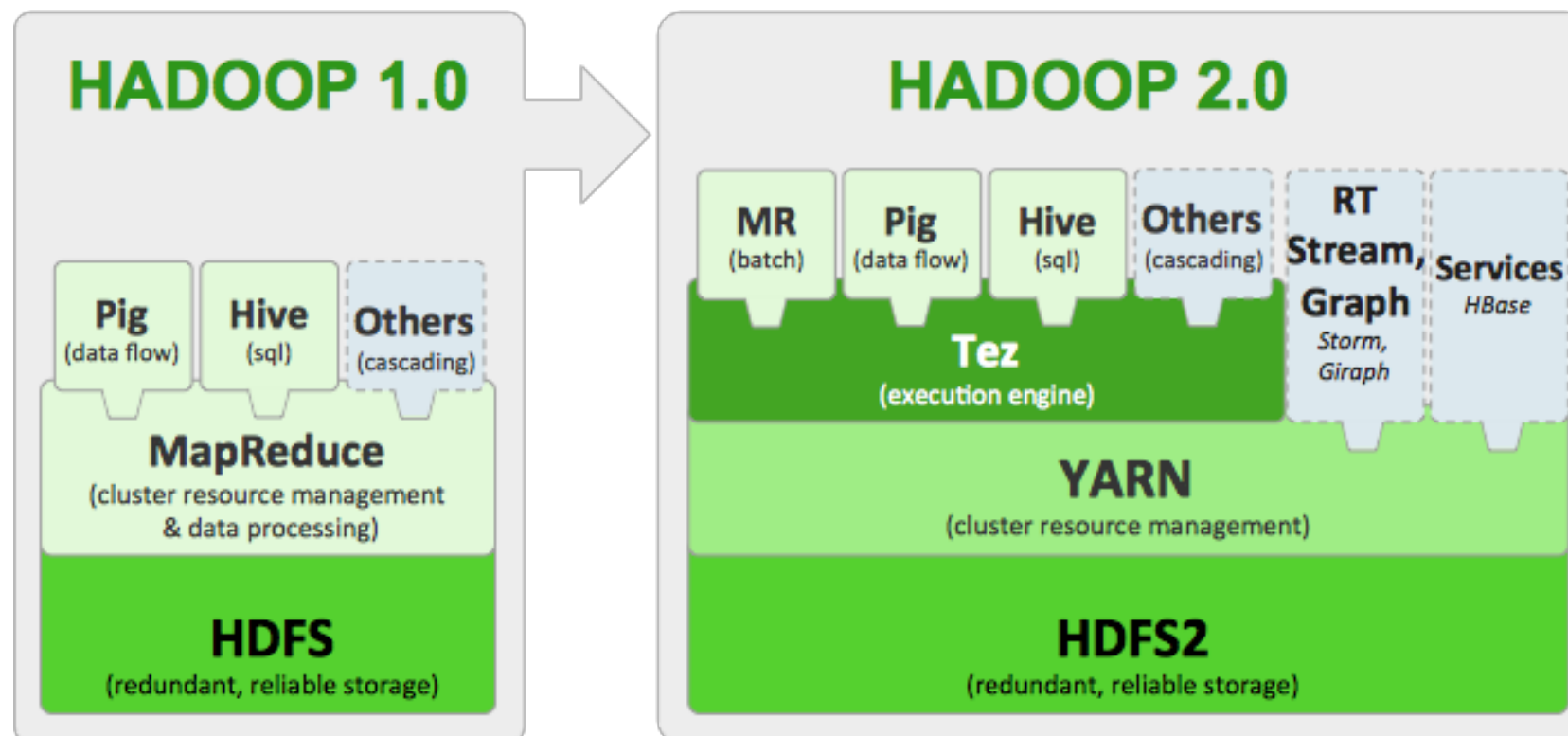


# Historia

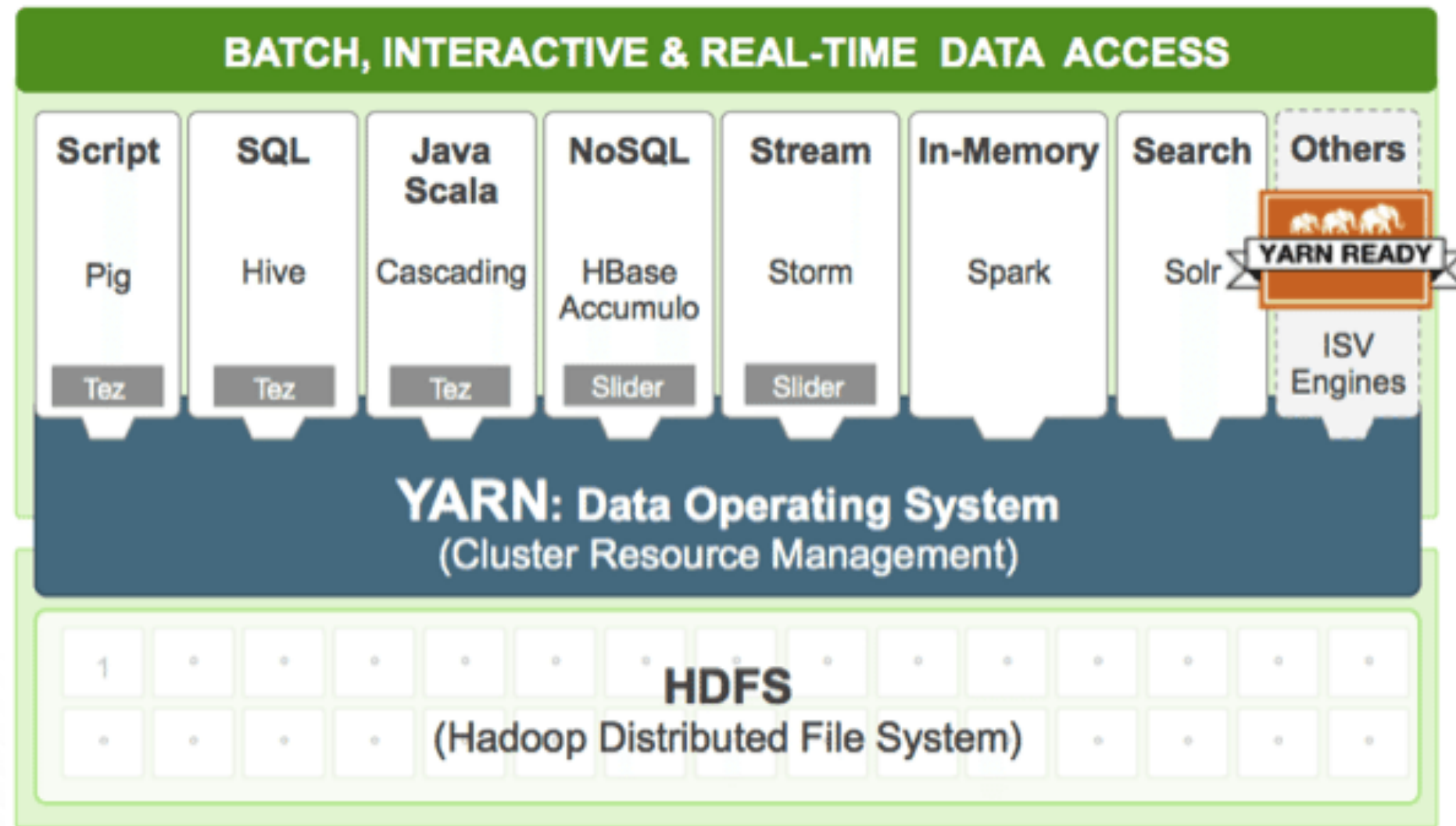




# Evolución

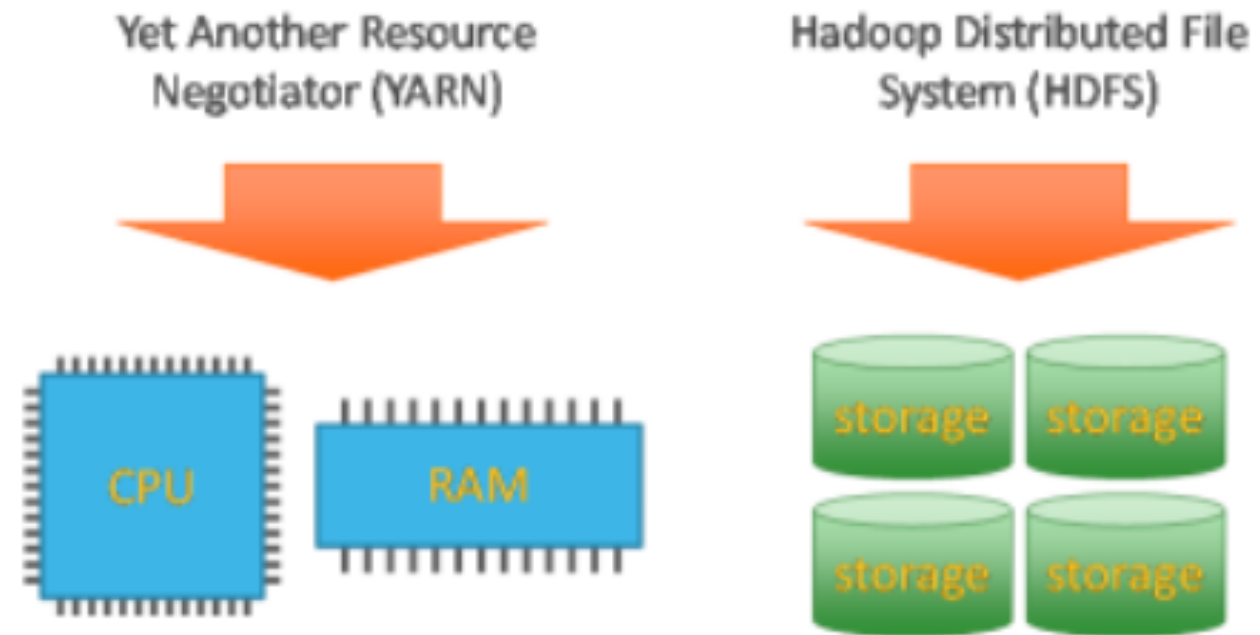


# Ecosistema

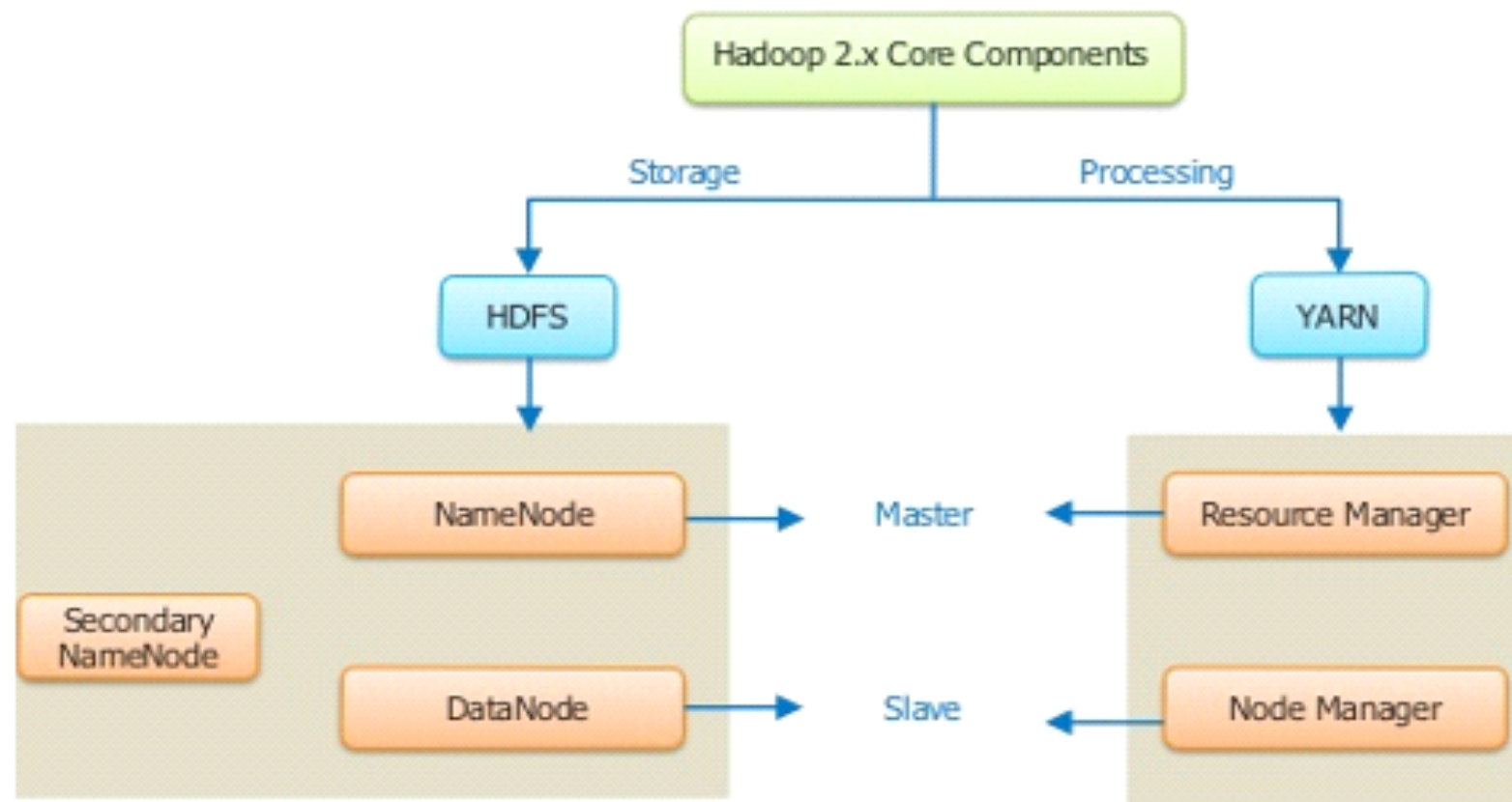


Apache Zookeeper

# Hadoop core

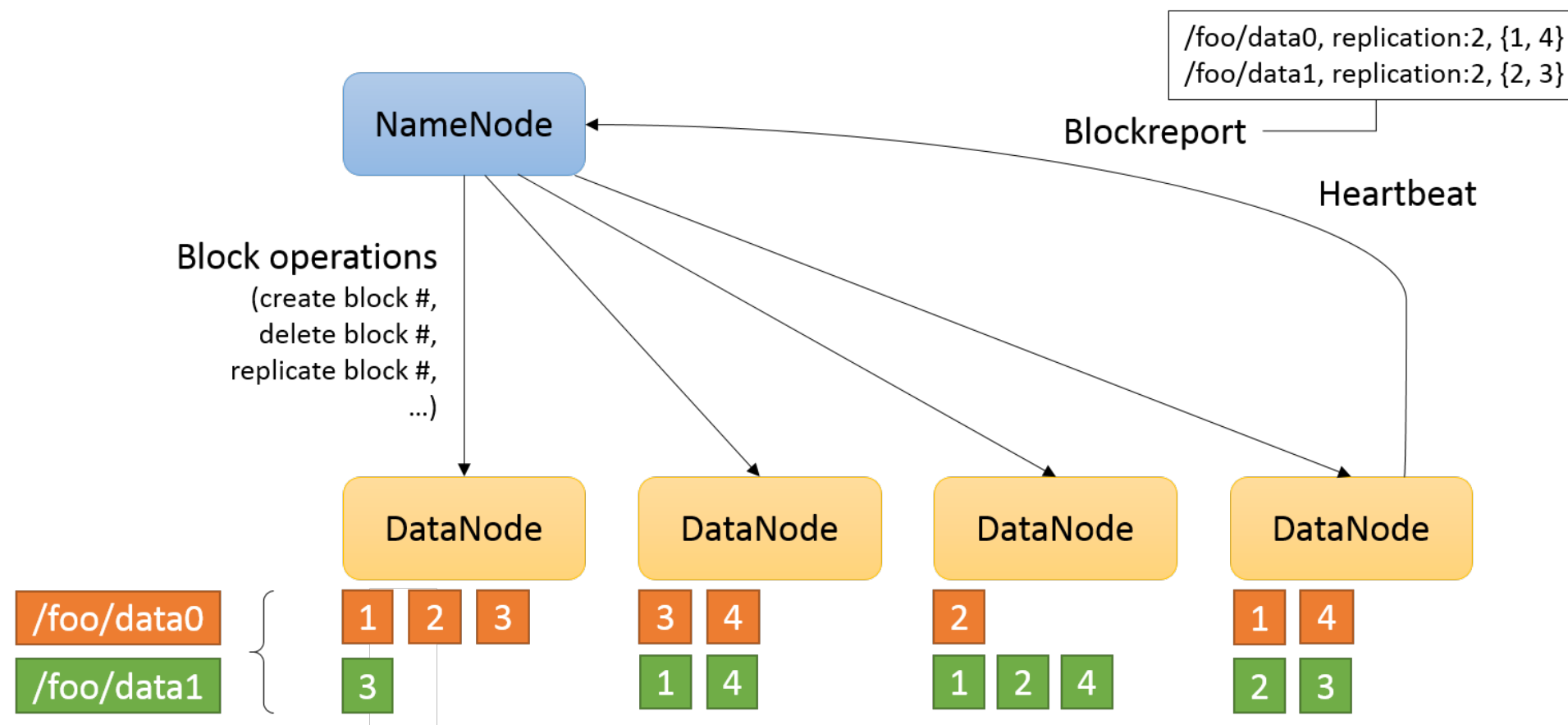


# Hadoop core



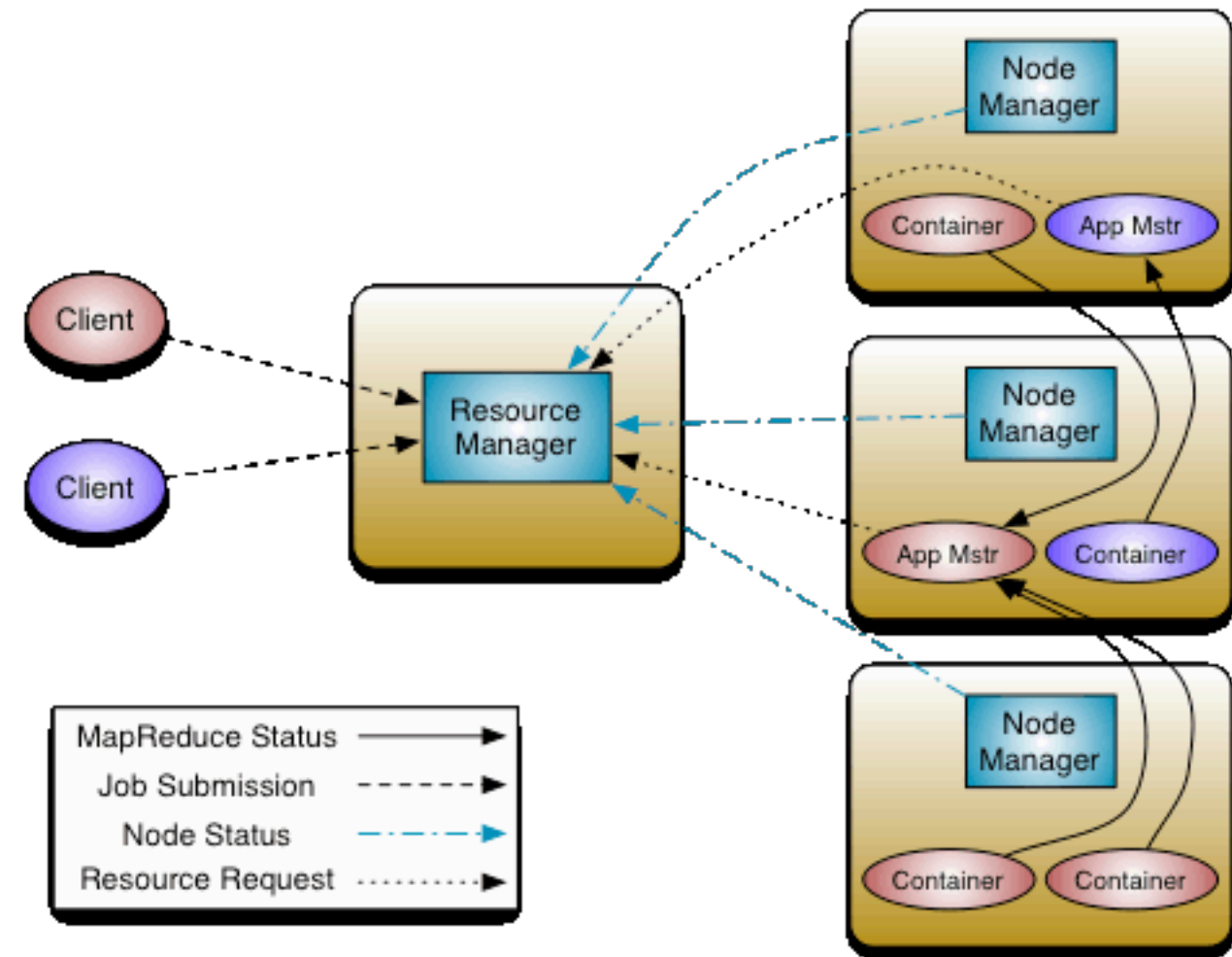
# HDFS (Hadoop Data File System)

- Namenode
- Datanode
- Client



# YARN (Yet Another Resource Negotiator)

- ◆ Client
- ◆ Resource Manager (Scheduler)
- ◆ Application Master
- ◆ Node Manager
- ◆ Resource ontainer





# Gracias!

