Jade Ullola Castro, 4500985

1 ECOSISTEMAS PARA GRANDES VOLÚMENES DE DATOS

ELASTICSEARCH

DEFINICIÓN

Motor de búsqueda y análisis de distribuido de código abierto. Diseñado para buscar, analizar y visualizar grandes volúmenes de datos de manera rápida y en tiempo real.

ANTECEDENTES

Creado por Shay Banon, 2010. Originalmente fue para limitaciones de motores de búsqueda existentes y proporcionar una solución que escale horizontalmente para manejar grandes cantidades de datos.

CARACTERÍSTICAS

Búsqueda de texto completo (estructurado o no)
Análisis en tiempo real
Escalabilidad horizontal (+nodos)
API RESTful
Utiliza el motor de indexación Apache Lucene.

ARQUITECTURA

Nodo. Almacenan datos y participa en las operaciones de indexación y búsqueda.
Índice. Colección lógica de documentos (c/entrada única y datos estructurados.
Clúster. Nodos que trabajan juntos y comparten la carga de trabajo y los datos.

COSTOS

Desde los 95 - 175 USD al mes

Jade Ullola Castro, 4500985

2 ECOSISTEMAS PARA GRANDES VOLÚMENES DE DATOS

APACHE SPARK

DEFINICIÓN

Framework de procesamiento de datos de código abierto y distribuido, con interfaz unificada para el procesamiento de datos en batch y en tiempo real.

ANTECEDENTES

Universidad de California, Berkeley (proyecto de investigación "AMPLab" (Berkeley Data Analytics Stack)). Proyecto de código abierto en 2010.

CARACTERÍSTICAS

Operaciones de procesamiento de datos en memoria
Proporciona una API unificada para el procesamiento de datos batch y streaming
Soporte para múltiples lenguajes
Librerías integradas
Puede escalar horizontalmente.

ARQUITECTURA

Driver Program. Lógica de la aplicación y contexto Spark.
Cluster Manager. Gestiona los recursos del clúster y programa tareas.
Cluster Worker Node. Ejecutan las tareas.
Spark Context. Coordina tareas y gestiona la comunicación.

COSTOS

Gratuito

Jade Ullola Castro, 4500985

3

ECOSISTEMAS PARA GRANDES VOLÚMENES DE DATOS

CASSANDRA

DEFINICIÓN

Sistema de gestión de bases de datos distribuidas y altamente escalable, diseñado para manejar grandes volúmenes de datos en entornos distribuidos.

ANTECEDENTES

1ero por Facebook

Posteriormente, proyecto de código abierto en 2008.

Luego, proyecto de nivel superior de la Apache Software Foundation.

CARACTERÍSTICAS

Crecer de manera lineal al agregar nodos adicionales al clúster.

Tolerancia a fallos

Almacena datos en modelo tipo clave-valor distribuido

Sin punto único de falla

Flexible y altamente configurable

ARQUITECTURA

Nodo. Almacena datos y hace operaciones de L/E.

Anillo (Ring). Nodos forman un anillo y c/u es responsable de un rango de datos en el anillo.

Espacio de claves (KeySpace). Unidad lógica que agrupa column families, define la replicación y las estrategias de consistencia.

Column Family. Tabla en una base de datos relacional y almacena datos.

COSTOS

Gratuito

Jade Ullola Castro, 4500985

4 ECOSISTEMAS PARA GRANDES VOLÚMENES DE DATOS

CEPH

DEFINICIÓN

Sistema de almacenamiento distribuido de código abierto diseñado para proporcionar almacenamiento escalable, de alto rendimiento y altamente tolerante a fallos.

ANTECEDENTES

Sage Weil, Universidad de California, Santa Cruz. Se lanzó como proyecto de código abierto y ha sido adoptado. Su desarrollo ha continuado bajo la dirección de la empresa Inktank (adquirida por Red Hat en 2014).

CARACTERÍSTICAS

Almacenamiento en bloques, archivos y objetos

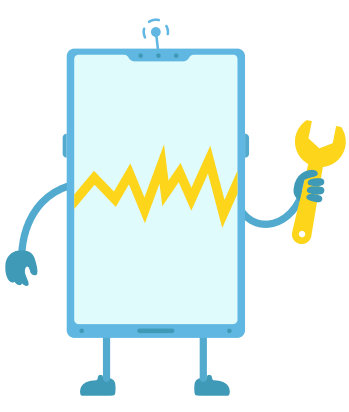
Escalabilidad horizontal

Tolerancia a fallos

Arquitectura modular

Interfaz RESTful para administración y monitorización del clúster.

ceph



ARQUITECTURA

RADOS (Replicated and Distributed Object Store). Capa de almacenamiento de objetos.

Ceph OSD Daemon (Object Storage Daemon). C/nodo en el clúster ejecuta OSDs.

Ceph Monitor Daemon. Monitorea el estado del clúster y coordina la configuración.

Ceph Metadata Server (MDS). Metadatos para CephFS.

COSTOS

Desde 0 - 14.99 USD al mes

Jade Ullola Castro, 4500985

5

ECOSISTEMAS PARA GRANDES VOLÚMENES DE DATOS

IBM SPECTRUM SCALE (GPFS)

DEFINICIÓN

Sistema de archivos paralelo y de objetos que proporciona un almacenamiento distribuido y altamente escalable.

ANTECEDENTES

GPFS desarrollado por IBM, 1990.
Cambio de nombre a IBM Spectrum Scale, 2015.

CARACTERÍSTICAS

Almacenamiento distribuido de datos en clústeres de servidores.
Paralelismo. Ejecución de operaciones en paralelo.
Soporte para diversos protocolos de acceso como NFS, SMB, FTP, y S3.
Funcionalidades avanzadas como instantáneas (snapshots) y replicación de datos para garantizar la disponibilidad y protección de los datos.

ARQUITECTURA

Nodos de almacenamiento. Almacena datos.
Nodos de acceso. Interfaces de acceso a los usuarios y aplicaciones mediante protocolos (NFS o SMB).
Utiliza una red de interconexión de alta velocidad para facilitar la comunicación y un rendimiento óptimo.

COSTOS

Desde 0 - 656 USD por terabyte

REF.

Instituto de Ingeniería del Conocimiento. 7 Herramientas Big Data para tu empresa. Recuperado de: https://www.iic.uam.es/innovacion/herramientas-big-data-para-empresa/
PowerData. (2014). ¿Existen alternativas a Hadoop? Recuperado de: https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/bid/397384/Existen-alternativas-a-Hadoop

Jade Ullola Castro, 4500985

SCREENSHOTS