PROGRAMA DE CIENCIAS DE LOS DATOS
CURSO: BIG DATA

PrestoDB

- **Tarea #2**

- **Profesor:** MSc. Felipe Meza

- **Alumno:**
- Lester Salazar Viales
- Randal Salazar Viales

- San José, 09 de diciembre de 2019.

PrestoDB

- **Historia**

- 2008: Facebook creó Apache Hive para ejecutar análisis SQL en su almacén de datos de múltiples petabytes en Hadoop.
- Problema:
  - Hive era inadecuado para la escala de Facebook y se inventó Presto (para ejecutar consultas rápidas).
- 2012: Comienza desarrollo de Presto con un sistema de consulta interactivo que podría operar rápidamente a escala de petabytes. Se lanza en primavera 2013 a toda la compañía.
- Noviembre 2013: Facebook abrió Presto bajo la Licencia de Software Apache (descargar en Github).

PrestoDB

- **Historia**

- 2014: Netflix reveló que usaron Presto en 10 petabytes de datos almacenados en el Amazon Simple Storage Service (S3).
- Enero 2019: se forma Presto Software Foundation organización sin fines de lucro dedicada al avance del motor de consulta SQL distribuido de código abierto de Presto).
  - **PrestoDB** : mantenido por Facebook (*https://prestodb.github.io.*)
  - **PrestoSQL** : mantenido por Presto Software Foundation con alguna polinización cruzada de código (*https://prestosql.io.*).
- Setiembre 2019: Facebook donó PrestoDB a la Fundación Linux que establece la Fundación Presto.

PrestoDB

- **¿ Qué es Presto ?**

- Es un motor de consulta basado en SQL que utiliza una arquitectura MPP (procesamiento masivo en paralelo) para escalar.

- Se basa en Java y también puede integrarse con otras fuentes de datos de terceros o componentes de infraestructura.

- Por ser un motor de consulta, separa el cómputo y el almacenamiento dependiendo de los conectores para integrarse con otras fuentes de datos para realizar consultas.

PrestoDB

- **¿ Qué es Presto ?**

- La capacidad de consultar contra:

  Bases de datos tradicionales
  MySQL
  PostgreSQL
  Microsoft SQL Server

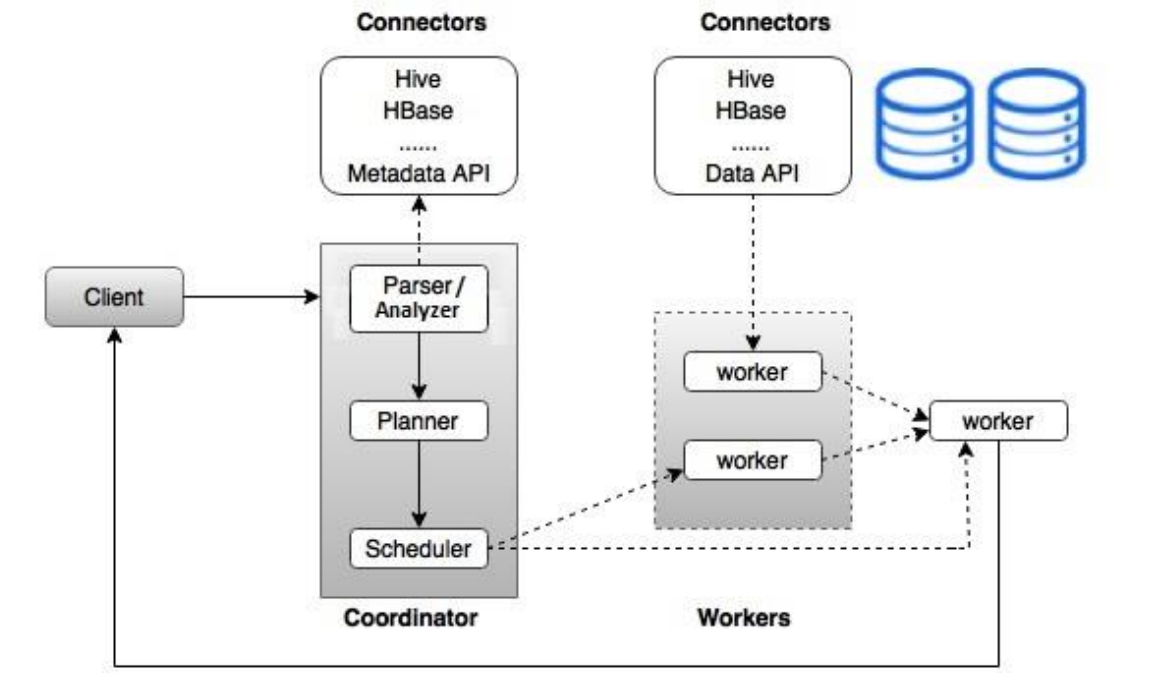
- Amazon Redshift
- Teradata
-

  Bases de datos no relacionales
  Mongodb
  Redis
  Cassandra

- Apache HBASE

Formatos de archivo en columnas como ORC, Parquet y Avro, almacenados en:
Amazon Simple Storage Service (Amazon S3)
Google Cloud Store
Tienda de blogs de Azure
Hadoop Distributed File System (HDFS)
Sistemas de archivos agrupados

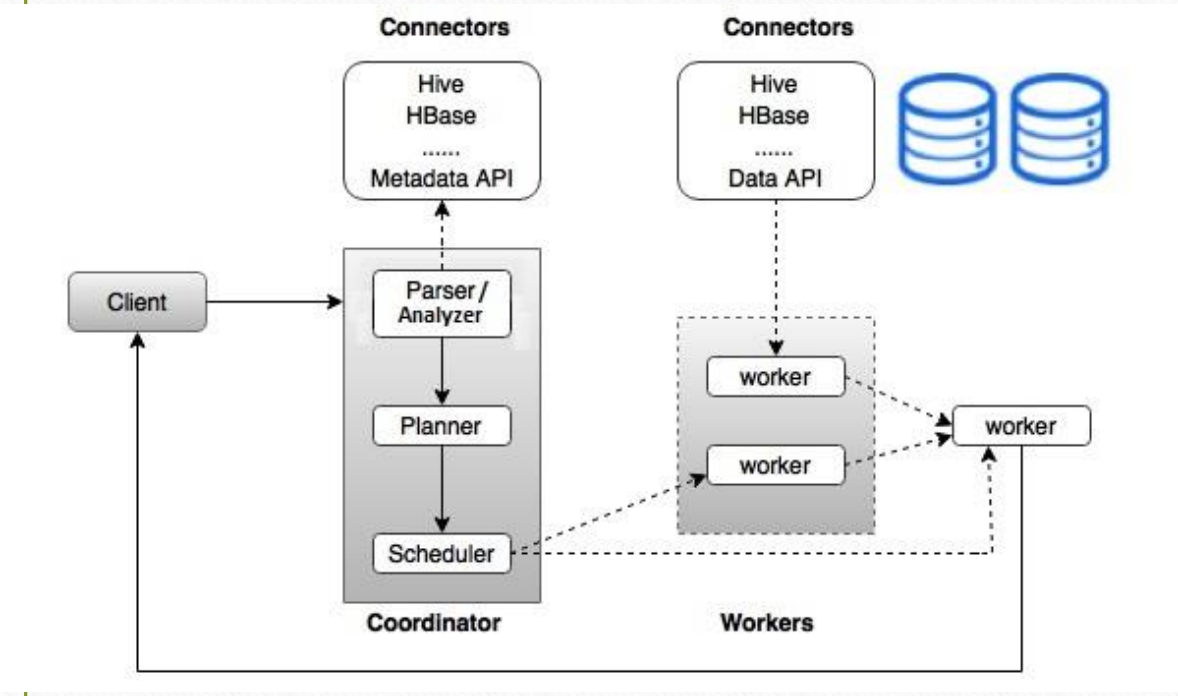
PrestoDB



- **Apache Presto - Architectura**

No.	Componente & Descripción
1	**Client** El cliente (Presto CLI) envía consultas SQL a un coordinador para obtener el resultado.
2	**Coordinator** El coordinador es un maestro. El coordinador analiza gramaticalmente inicialmente las consultas SQL y luego analiza y planifica la ejecución de la consulta. El programador realiza la ejecución de la canalización, asigna el trabajo al nodo más cercano y supervisa el progreso.
3	**Connector** Los complementos de almacenamiento se llaman conectores. Hive, HBase, MySQL, Cassandra y muchos más actúan como un conector; de lo contrario, también puede implementar uno personalizado. El conector proporciona metadatos y datos para consultas. El coordinador usa el conector para obtener metadatos para construir un plan de consulta.
4	**Worker** El coordinador asigna tareas a los nodos de trabajo. Los trabajadores obtienen datos reales del conector. Finalmente, el nodo trabajador entrega el resultado al cliente.

PrestoDB

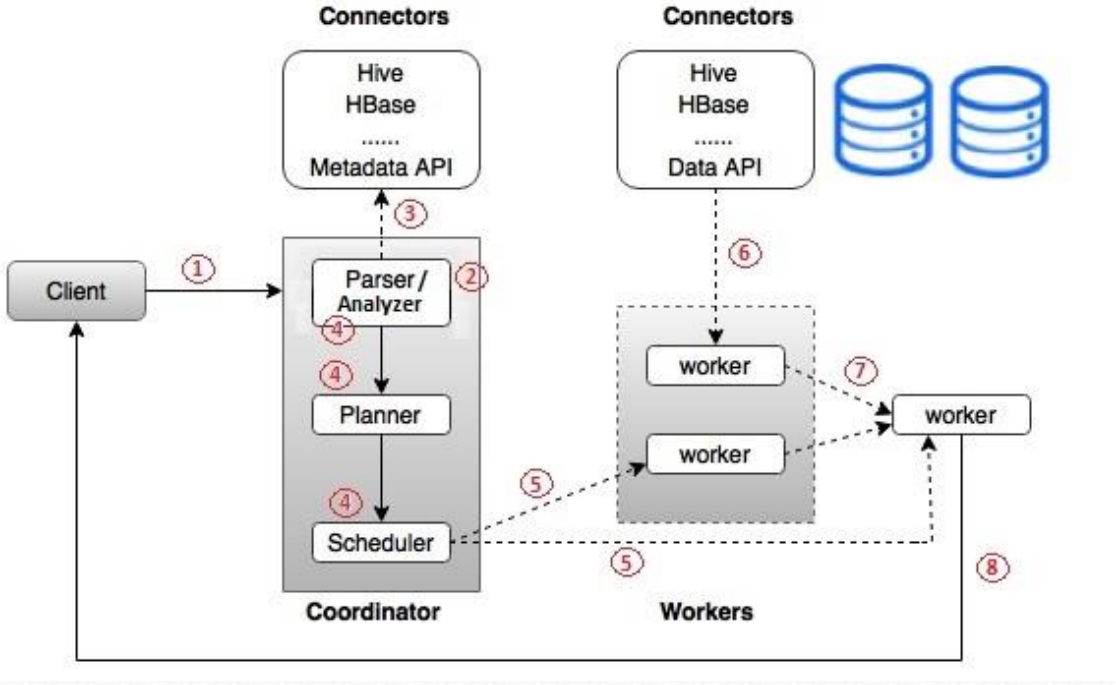


- **Presto - Flujo de trabajo**

- Presto es un sistema distribuido que se ejecuta en un cluster de nodos y utiliza una arquitectura similar a un sistema clásico de gestión de bases de datos de procesamiento masivo en paralelo (MPP).

- Tiene un nodo coordinador que trabaja en sincronización con múltiples nodos trabajadores.

PrestoDB

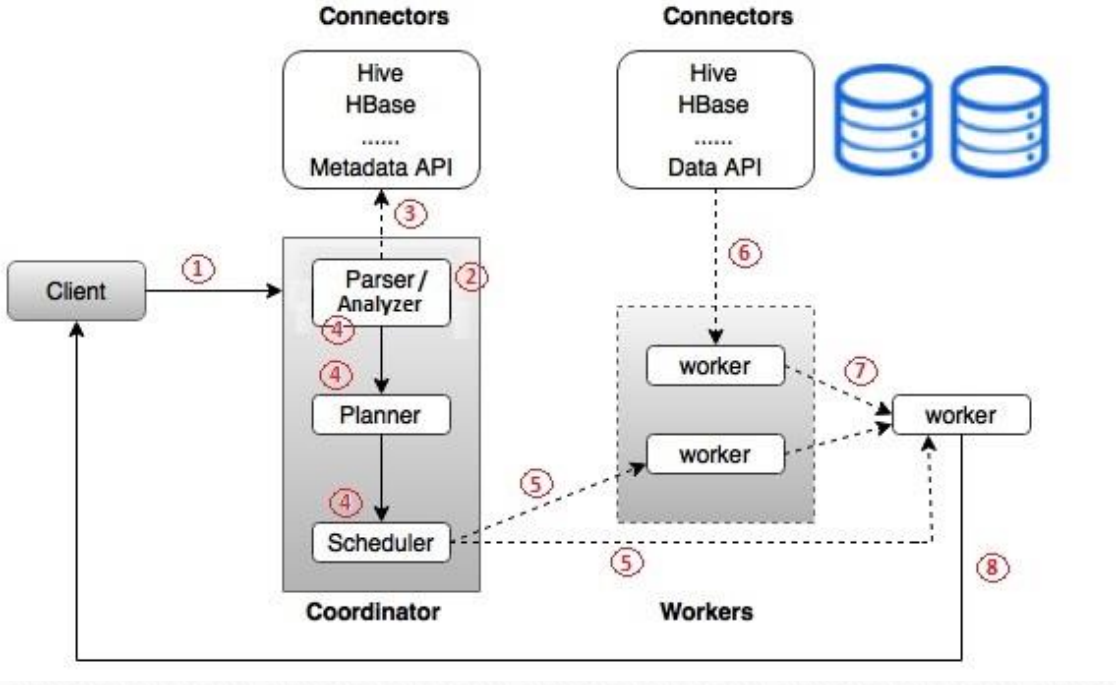


- **Presto - Flujo de trabajo**

1) El cliente (Presto CLI – **Client** ó usuarios) envían su consulta SQL al coordinador.

2) El coordinador inicialmente analiza gramaticalmente las consultas SQL (instrucciones) enviadas. Está diseñado para admitir la semántica ANSI SQL estándar, incluidas consultas complejas, agregaciones, combinaciones, combinaciones externas izquierda / derecha, subconsultas, funciones de ventana, recuentos distintos y percentiles aproximados.

PrestoDB

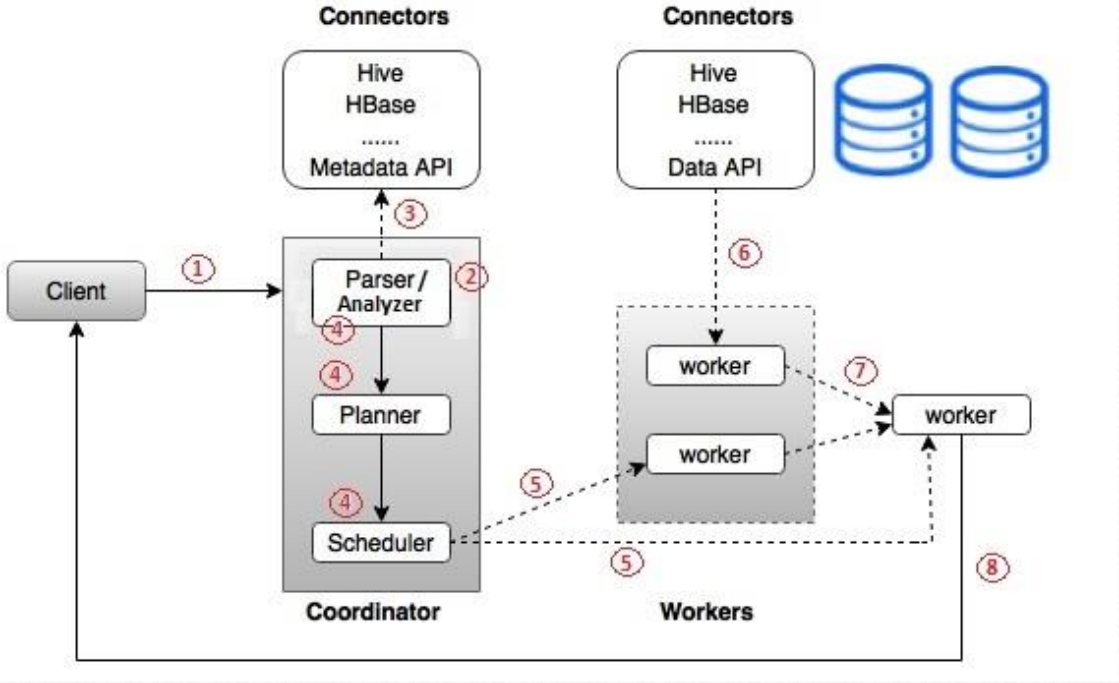


- **Presto - Flujo de trabajo**

3) Con la consulta comprendida, el coordinador utiliza el **conector** (storage pluggin) que está siendo empleado para obtener metadatos (los datos que describen otros datos). De esta forma, el coordinador conoce dónde localizar los datos que se quieren accesar. El coordinador usa el conector para obtener metadatos para construir un plan de consulta.

4) El coordinador utiliza un motor de consulta y ejecución personalizado para **analizar**, **planificar** y **programar** un plan de consulta distribuido en los nodos trabajadores.

PrestoDB



- **Presto - Flujo de trabajo**

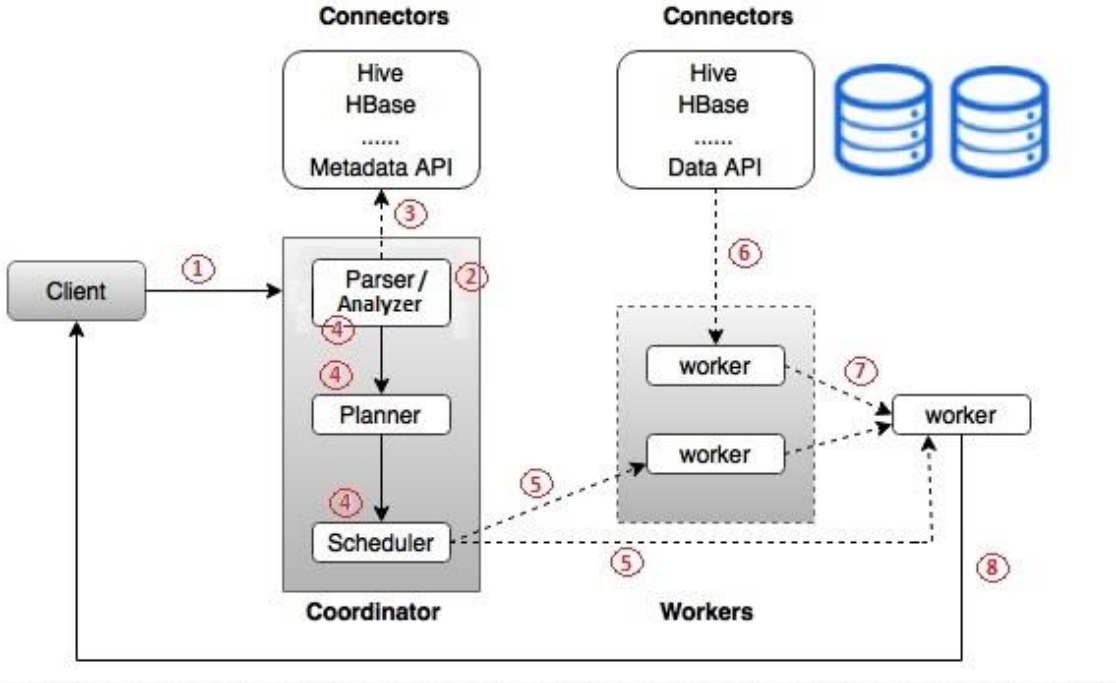
- 5) El planificador asigna:

  - trabajo a los nodos más cercanos a los datos (**nodos trabajadores**) y supervisa el progreso.

  - designa el **nodo trabajador** (encargado de brindarle los datos al cliente).

- 6) Los nodos trabajadores obtienen los datos solicitados mediante el conector empleado para extraer los datos (Data API).

PrestoDB



- **Presto - Flujo de trabajo**

- 7) Los nodos trabajadores se encargan de trasladar la información extraída al nodo trabajador.

- 8) Finalmente, el nodo trabajador procesa los datos que le fueron facilitados por los nodos trabajadores y entrega el resultado al cliente. El cliente extrae datos del proceso de salida.

PrestoDB

- **Modelo de ejecución**

- Los datos se consultan donde se almacenan, sin la necesidad de moverlos a un sistema de análisis separado.

- Además de la programación mejorada, todo el procesamiento está en la memoria y se canaliza a través de la red entre diferentes etapas.

- Esto evita la sobrecarga innecesaria de latencia de E / S.

PrestoDB

- **¿ Quiénes usan Presto ?**

La implementación de Presto en Facebook es utilizada por más de mil empleados, que ejecutan más de 30,000 consultas, procesando un petabyte de datos diariamente.

En promedio, Netflix ejecuta alrededor de 3,500 consultas por día en sus clústeres Presto.

PrestoDB

- **Referencias**

- https://en.wikipedia.org/wiki/Presto_(SQL_query_engine)
- https://aws.amazon.com/es/big-data/what-is-presto/
- https://www.tutorialspoint.com/apache_presto/apache_presto_architecture.htm
- https://www.alluxio.io/learn/presto/
- https://blog.openbridge.com/what-is-facebook-presto-presto-database-or-prestodb-a-powerful-sql-query-engine-77d4c4a66d4
- https://optimalbi.com/blog/2018/05/15/setting-up-prestodb-on-linux/

- **Link MEDIUM (**artículo PrestoDB en Medium**):**

- https://medium.com/@r.salazarvi/prestodb-22c84eca202

Gracias por su atención