Survey of distributed system based OLAP

Hao wang, Wang xi

Introduction

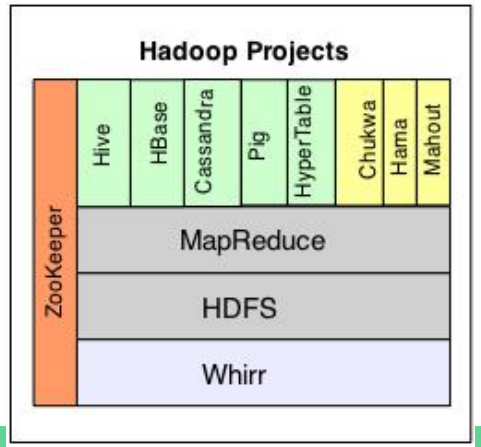
Distributed Database: Concept, MapReduce, Hadoop, Cutting-edge applications

OLAP: Concept, types, MR-Cube

Distributed Database

A **distributed database** is a database in which storage devices are not all attached to a common processor. It may be stored in multiple computers, located in the same physical location; or may be dispersed over a network of interconnected computers. Unlike parallel systems, in which the processors are tightly coupled and constitute a single database system, a distributed database system consists of loosely coupled sites that share no physical components. [via Wikipedia]

MapReduce Introduction ➔ Hadoop project



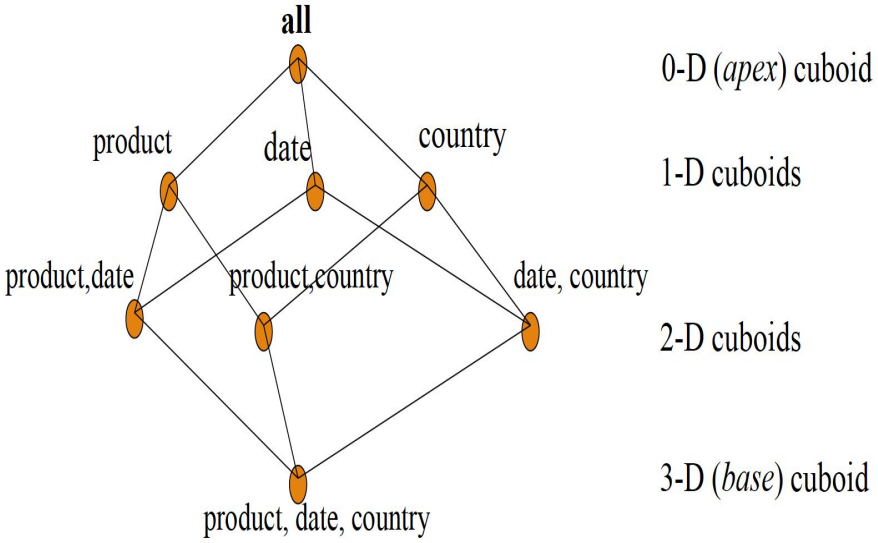
OLAP (data cube)

Dimension: a set of data features
eg. product, date, country

Level: different aspects of dimension
eg. Date: year, month, day

Member: member of specific dimension
eg. all products data in Apri 13 at Urbana

Measure: basic unit of OLAP cube
eg. sale data of shampoo in Apri 13 at Urbana



OLAP (operation)

Roll up (drill-up): summarize data

climbing up hierarchy or by dimension reduction

Drill down (roll down): reverse of roll-up

higher level summary → lower level summary or detailed data

introducing new dimensions to gather data we interested in

Slice and dice: project and select

Pivot (rotate):

reorient the cube to translate 3D → 2D planes

OLAP (types)

Multidimensional OLAP (MOLAP)

stores data in optimized multi-dimensional arrays storage

Relational OLAP (ROLAP)

stores data in relational database

Hybrid OLAP (HOLAP)

combination of MOLAP and ROLAP

MR-cube(building)

Full Source Scan:

Using HBase facilities to scan the whole source filtering it by the attributes the user

Indexed Random Access:

Building indexes beforehand to easily obtain the identifiers of the desired tuples and then retrieve the data by random access

Index Filtered Scan:

Combination of above two approaches

MR-cube(computation)

Partially algebraic measures:

Computing from sub-groups:

(1)mutually exclusive on the full tuple

(2)mutually exclusive after projecting on the algebraic attribute

Sampling approach:

Generating sample from cube computation on small random dataset

According to the result of sample, divide date into reducer-friendly and reducer unfriendly parts

Batch areas:

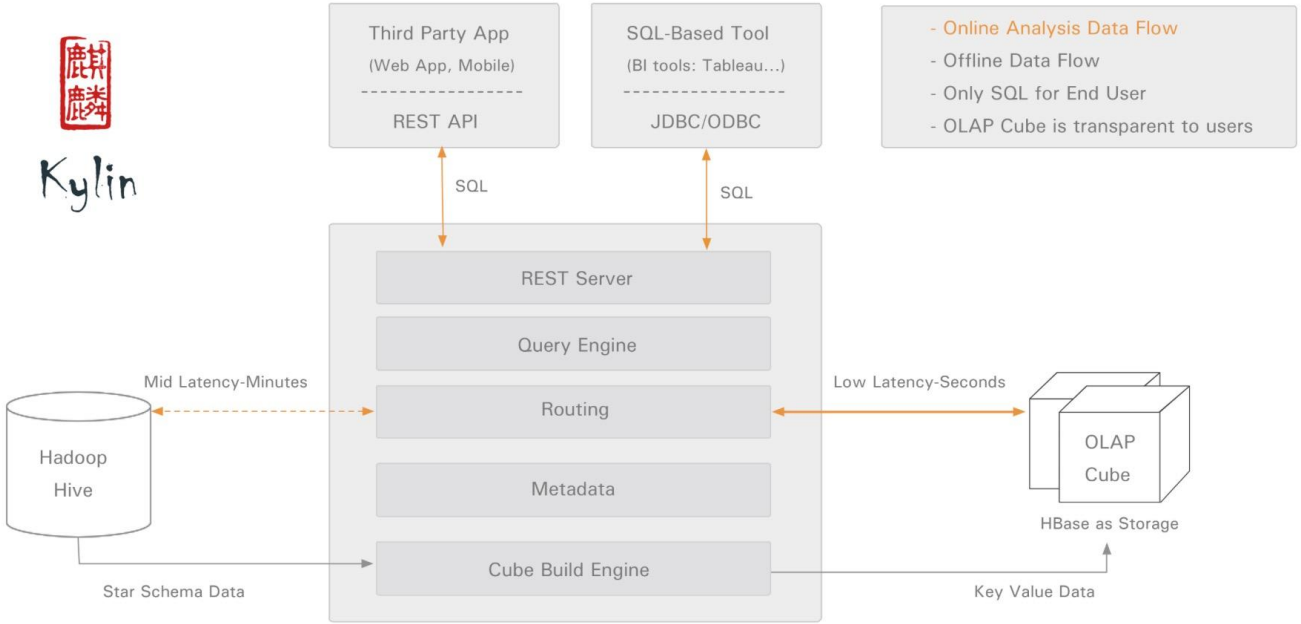
Map: emits one key-value pair per batch for each data tuple

→ reducing the amount of intermediate data

Reduce: executes traditional cube computation algorithm over results of map step

Application

Apache Kylin



Conclusion

The Distributed System’s development

The OLAP technology’s development

Apache Kylin project applications on OLAP based Distributed System

Reference

Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. Communications of the ACM, 51(1):107–113, 2008.
RayonStorage blog. Hadoop introduction. 2011
Jeffrey Dean and Sanjay Ghemawat. Mapreduce: a flexible data processing tool. Communications of the ACM, 53(1):72– 77, 2010.
J. Dean, S. Ghemawat, Mapreduce: simplified data processing on large clusters, Communications of the ACM 51 (1) (2008) 107–113.
W. H. Inmon, Building the data warehouse, John wiley & sons, 2005.
E. F. Codd, S. B. Codd, C. T. Salley, Providing olap (on-line analytical 135 processing) to user-analysts: An it mandate, Codd and Date 32.
S. Chaudhuri, U. Dayal, An overview of data warehousing and olap technology, ACM Sigmod record 26 (1) (1997) 65–74.
O. Council, Olap and olap server definitions (1997).
A. Abell´o, J. Ferrarons, O. Romero, Building cubes with mapreduce, in: 140 Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP, ACM, 2011, pp. 17–24.
A. Nandi, C. Yu, P. Bohannon, R. Ramakrishnan, Data cube materialization and mining over mapreduce, IEEE transactions on knowledge and data engineering 24 (10) (2012) 1747–1759.