RESEARCH DETAIL

BIG DATA CONTROL PLATFORM IN TIMES OF EPIDEMIC.

An application of big data technology in epidemic prevention and control.

PROJECT OVERVIEW

- The project discusses the use of a Hadoop architecture to store large amounts of subscriber mobile phone signalling data to support close contact tracing during an outbreak.
- The project uses the architecture of a Spark cluster and Hadoop, including Master and Worker nodes, and their roles in task scheduling and execution.
- This project describes a process for processing mobile phone signalling data using Spark, including the use of operations such as map, filter, cache and reduce to identify close contacts.

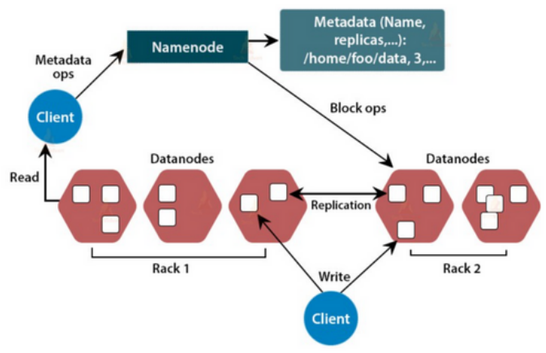
PROJECT BACKGROUND

- Big data technologies have a wide range of applications in medical and public health fields, such as disease diagnosis, prediction, gene sequence identification and health management. During the COVID-19 epidemic, big data enabled rapid epidemic prevention and precise epidemic control, and the information system provided important support for epidemic prevention and control by collecting, processing, analysing, mining and visual representation of data.

PROBLEM TO SOLVE

- How to efficiently store and manage massive mobile phone signalling data generated by the system in Hadoop architecture?
- How to efficiently process and analyze mobile phone signalling data to identify close contacts using MapReduce programs?
- How to implement efficient data parallel processing in Spark architecture?
- How to choose the right clustering algorithm to determine the best location for nucleic acid testing centers?

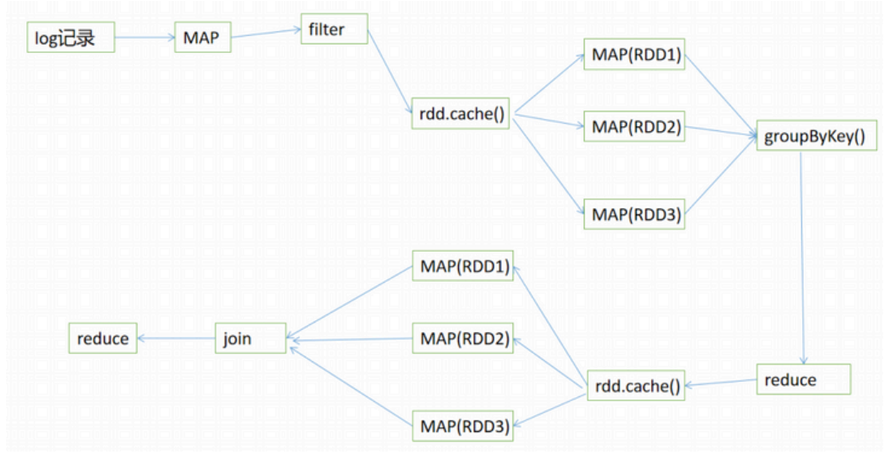
HADOOP STORAGE DETAILS



In HDFS, files are divided into blocks, and file access follows single-write and multi-read semantics. To satisfy fault tolerance requirements, multiple copies of a block are stored on different DataNodes

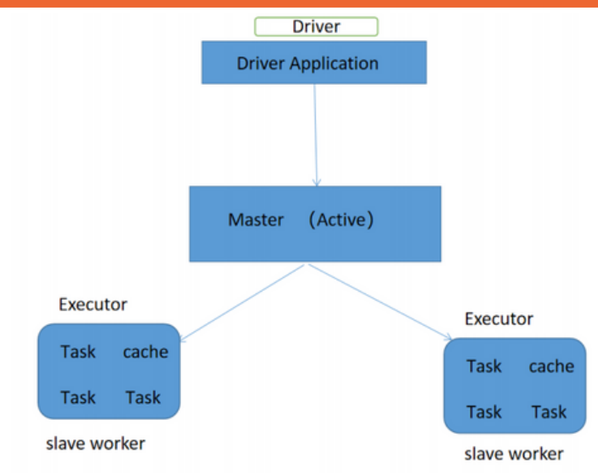
SPARK QUERY CATCHER

Each subscriber mobile phone signalling record will first go through MAP to produce a new RDD, which consists of each input element converted by the func function, and then these RDDs will go through a filter, which consists of input elements with a return value of true after being calculated by the func function,



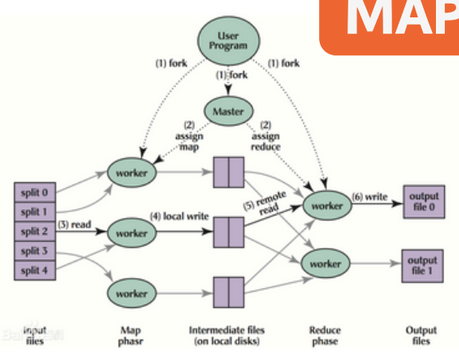
so that the RDDs that participate in the subsequent calculations will be filtered. This can play a filtering role, so that the signal records of the RDDs involved in the subsequent calculation are all complete, and then a RDD caching instruction is added to avoid repeated calculations and thus reduce the time.

SPARK SYSTEM ARCHITECTURE



In a Spark cluster, there is a master node and a worker node, where the master daemon and the driver process reside on the master node. The master is responsible for converting serial tasks into parallel executable Tasks, as well as error handling, etc., while the worker node has a master daemon and a driver process. The Worker node hosts the Worker daemon

MAPREDUCE POCESS MODEL

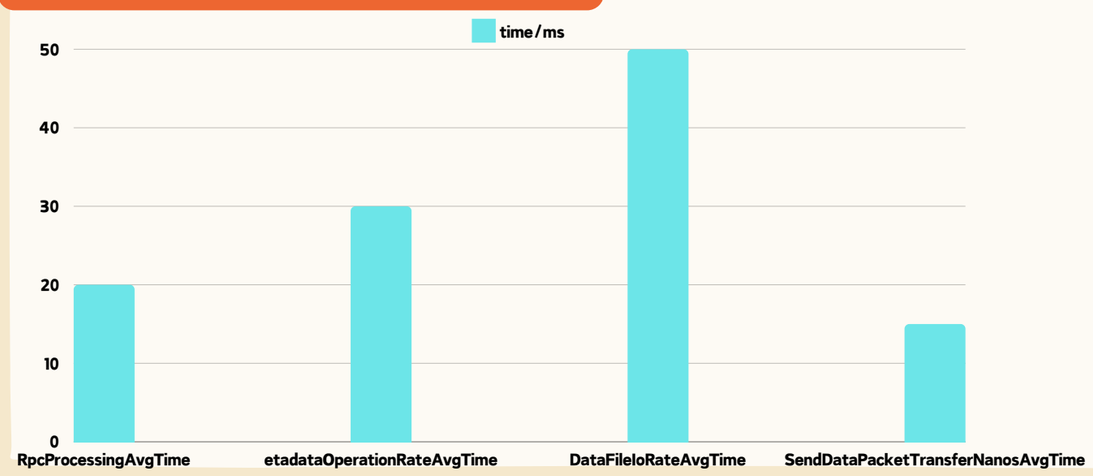


Automatically assigning and executing tasks and collecting results on cluster nodes, and parallelising computations such as data distribution and storage, data communication, and fault-tolerant processing.

CITATION

1.Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. Communications of the ACM, 51(1), 107-113.
2.Gu, X., Shang, Z., & Guo, X. (2020). A preliminary analysis of the application of information systems in the COVID-19 pandemic under the background of big data. Computer Science and Application, 10(12), 2197-2204.

PROJECT OUTCOMES



CONCLUSION

- The project discusses the application of big data in various industries, especially its important role during epidemics.
- The close relationship between big data and cloud computing is emphasised, as well as the potential of big data in government management.
- The projects points out that big data technology is still in the developmental stage, but its application is promising and industries should take advantage of this trend.

CONTACT ME :19302010047@FUDAN.EDU.CN