Volume Management of Hadoop DataNode

Xu Wang, Leitao Guo, and **Zhiguo Luo**
China Mobile Research Institute(CMRI)
June 1, 2011

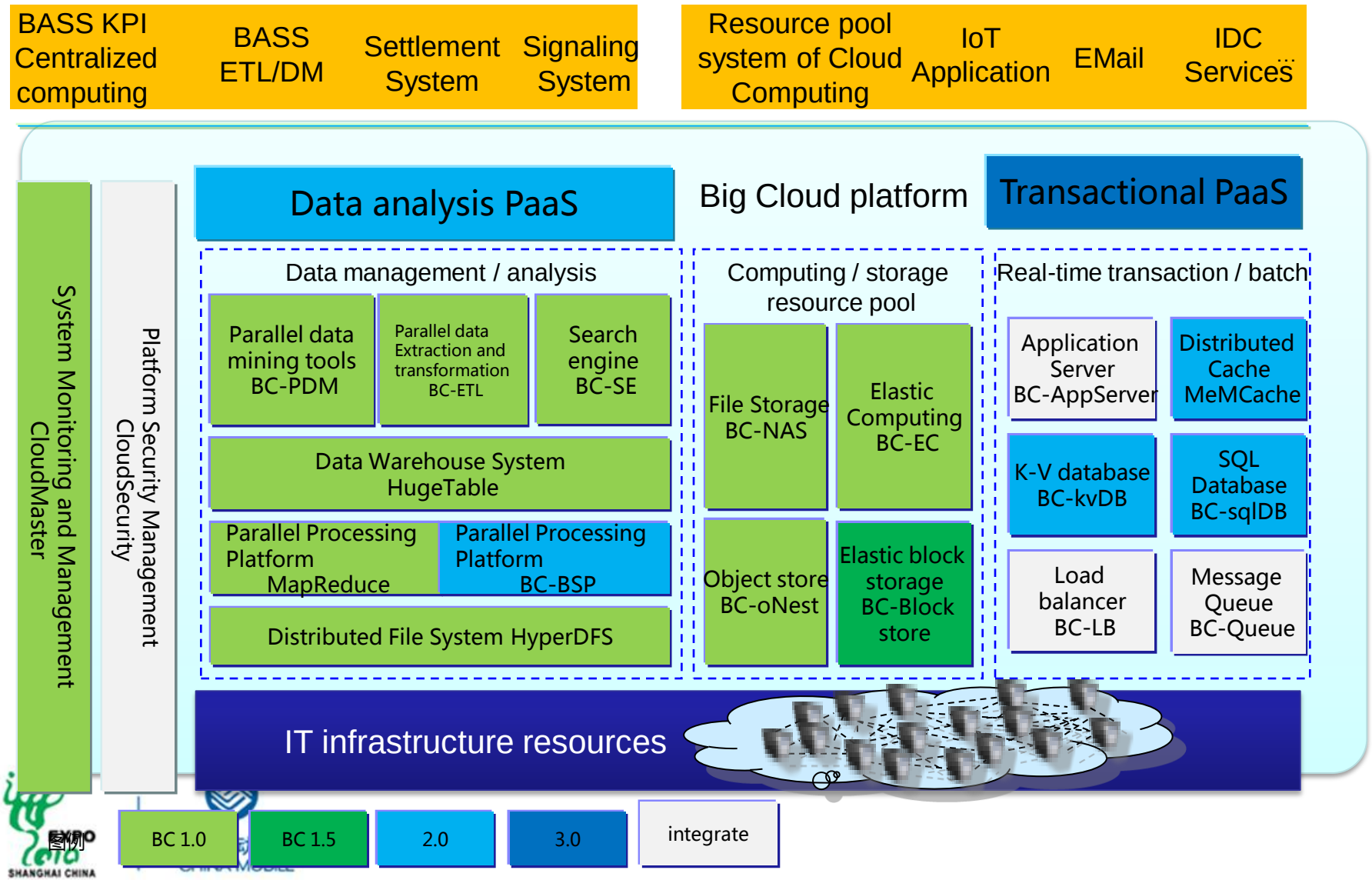
Agenda

- Cloud R&D in China Mobile
- Volume Management of Hadoop DataNode

CMRI “Big Cloud” Project

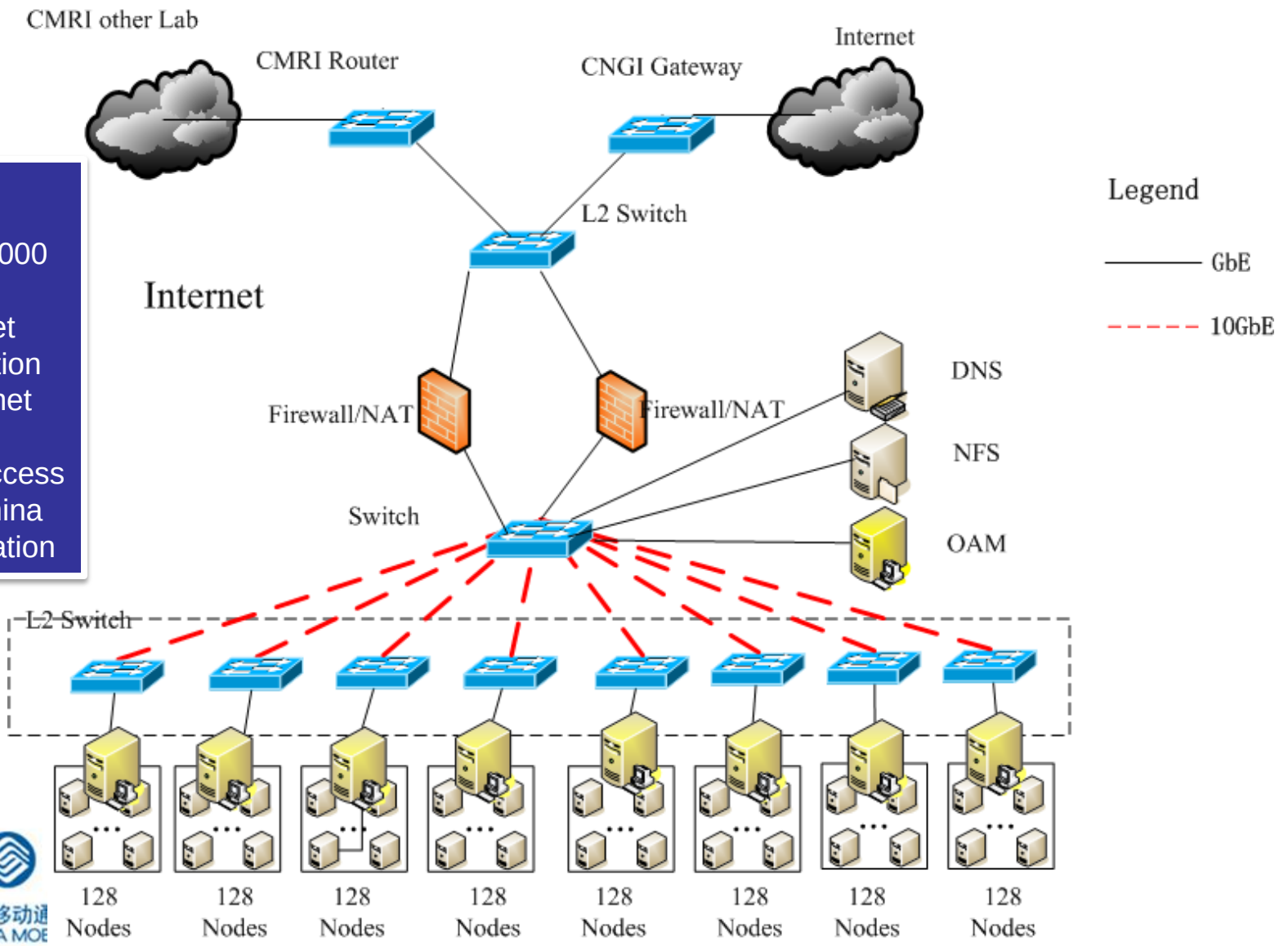
- What is “Big Cloud”
  - A CMRI project: research of key technologies and prototype system development for China Mobile Cloud Computing Infrastructure
- Goal
  - To meet the growing demands for high performance, low cost, high scalability, high reliability of China Mobile IT Infrastructure (computing, storage)
  - To meet the demands of China Mobile to deliver Internet business and services

“Big Cloud " Product Evolution Direction



Testbed in CMRI Lab

- 1024 Nodes More than 5000 SATA disks
- 1Gb Ethernet Interconnection
- 10Gb Ethernet Inter cluster
- 155Mbps Access to CNGI (China Next Generation



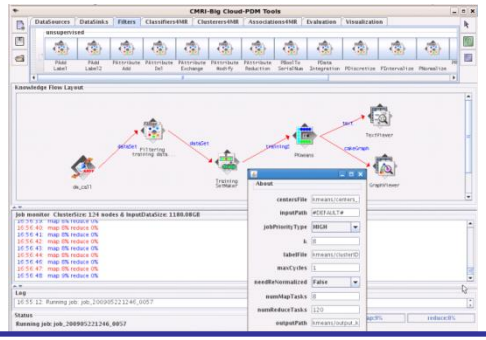
Cloud based Parallel Data Mining tool——BC-PDM/ETL

- A data mining solution for large-scale data analysis
  - Massive scalability - based on Hadoop
  - Low cost - commodity machines and open source software
  - Customization – facing to application requirements
  - Easy to use - similar user interface to commercial tools

- Parallel Data Mining algorithms based on MapReduce
  - ETL operations – parallel data extract, clean/transfer, load operations
  - Data mining algorithms – parallel classification, clustering, association rules algorithms
  - Data Exploration – parallel basic statistics descriptor, univariable analysis, multivariable analysis algorithms
  - Social Network Analysis – parallel network structure analysis, social network discovery, social network development algorithms

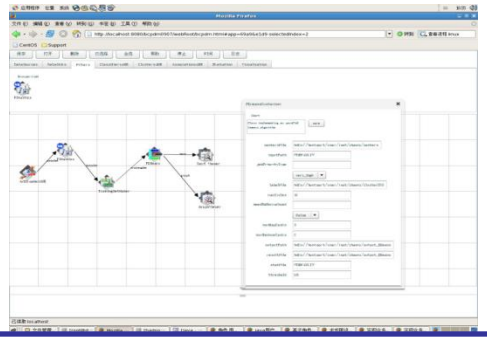
BC-PDM/ETL Evolution

Version I (2008)



- Workflow management
  - GUI - Drag Operation for application modeling design
  - Job Monitoring
  - Flow Configuration
- ETL (14 different ETL operations from 6 categories)
  - Statistic, attribute processing, data sampling, query, data processing, redundancy data processing
- Data mining Algorithm (9 algorithms from 3 categories)
  - Clustering, Classifier, Association Analysis
- Visualization
  - Text, decision tree, cake graph, and histogram

Version II (2009-2010)

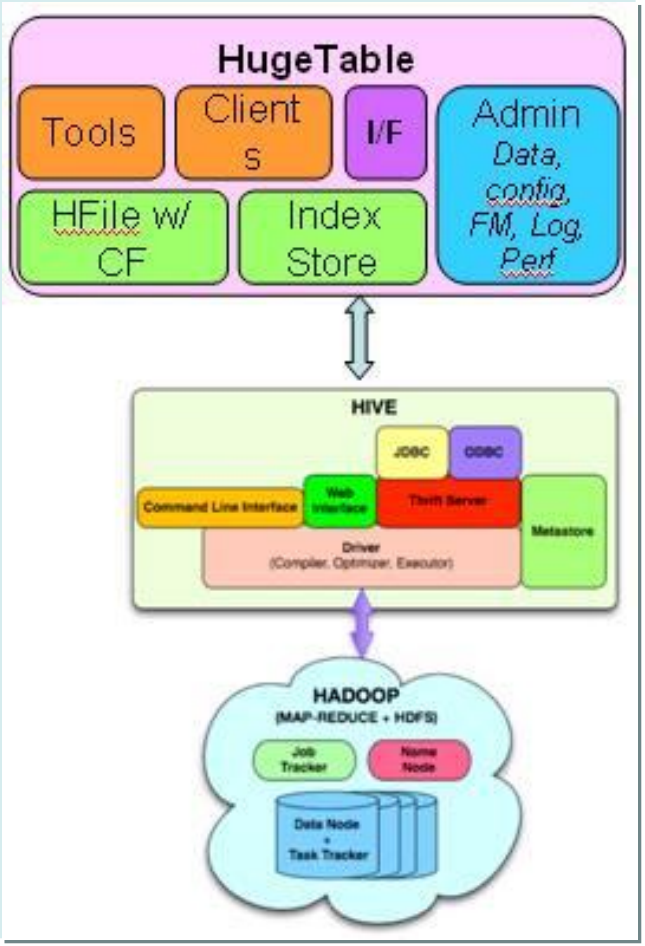


- Web based GUI: SaaS mode for users
- DE(Data Exploration): Simple data analysis and preview
- ETL (25 more): Join, Group by, Expression, case when, Update, and etc.
- Data mining Algorithm (4 more): Classifier, Sequence Association Analysis
- Data Transfer Tool: Provide data upload and download tools for SaaS
- Security: Multi-talent and user group for branch, ACL for data access

Version III (2010-2011)

- Web based GUI
  - SaaS mode for users
- UI: Workflow GUI, CLI, SQL
- Data Exchange: between DB and HDFS
- Integrate Hive: SQL-like scripts
- Parallel Social Network Analysis
- ETL(45 from 7 categories) DM(14 from 3 categories) Data exploration (3 clategories)
- application schedule system
- Application templates
- API: java Api, web service

Hugetable：Structured Data Management System and Applications



- Build on top of HDFS、MapReduce、Hive 、HBase and Zookeeper
- Provide index, join, fast load/export
- Provide JDBC/SQL, MapReduce, native API under access control
- Open Management Interface for system management
- Some system integrator companies are integrating the Hugetable into their applications for CMCC

EXPO 2010 SHANGHAI CHINA

中国移动通信 CHINA MOBILE

Agenda

- Cloud R&D in China Mobile
- Volume Management of Hadoop DataNode

Background

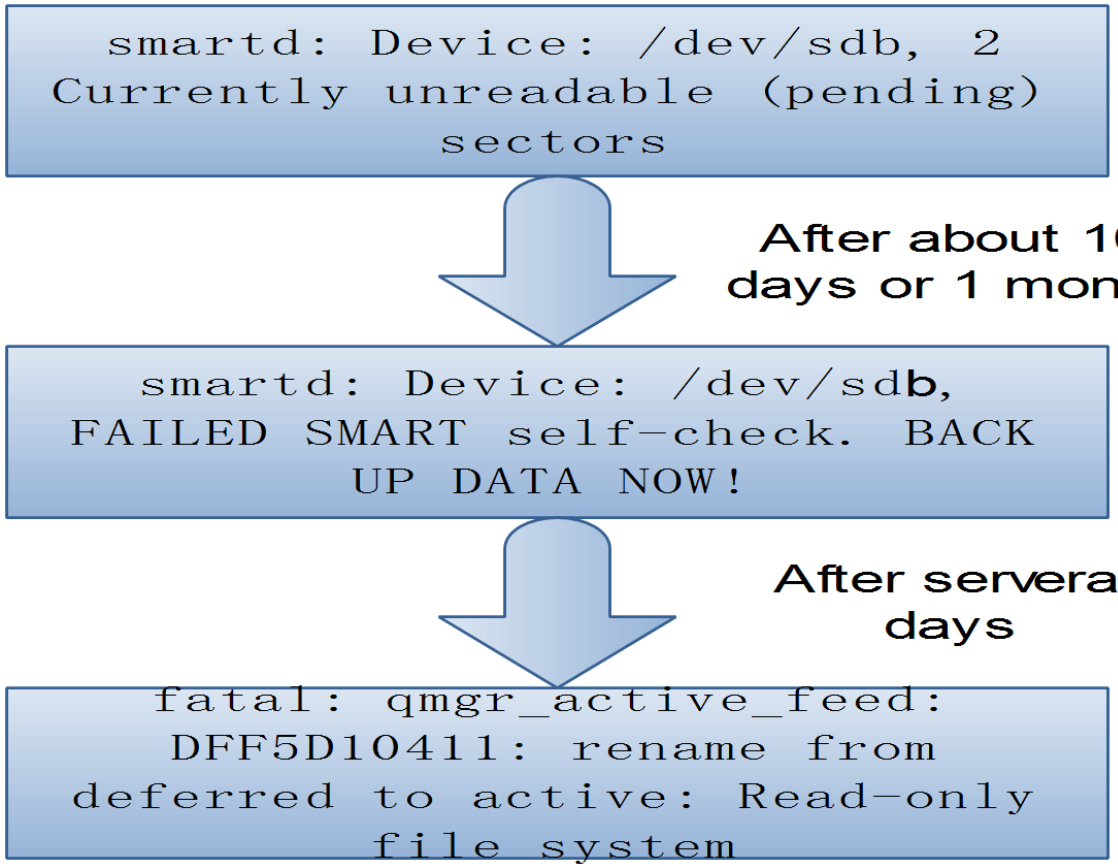
SATA Disk failure happens much more frequently than other failure

Servers tend to equip with more disks

→ Servers should be able to survive through disk failures

- HDFS 0.20 and before: disk fail is a fatal error of Hadoop Datanode
- HDFS-457: disk fail will lose a disk, but the server can still work.

Hard disk failure observation



We can guess the disks' health status from smartd information, find the sub-health disks and then take action ahead of time.

Objectives

- To forecast the disk failure automatically, and take actions ahead of time to improve the data availability, such as
  - to notify the NameNode to replicate the blocks stored in the failure disk by DataNode block-report. [similar with HDFS-457]
  - or to move the blocks to other healthy disks in the same DataNode and then invalidate the sub-healthy disk. [invalidate a disk online]

Implementation status

- Hadoop community don’t care about forecasting the disk failure too much, but regard it’s useful to replace the failure disk online.
- So, we implemented the functionality of changing disks online (HDFS-1362) firstly.
- We also tried to implement a functionality of invalidating a disk online but not open source until now.
- In our testbed, we automatically observe and guess the disk status, and trigger the action externally.

Volume Management of Datanode

- FSVolumeSet
  - Maintain volumes:
    - Scan volumes while startup
    - Remove failed volume (feature of HDFS-457)
- ReplicaMap
  - Maintain blocks
    - Map block to on-disk file
    - Responsible for all kinds of block related operation
  - You need traverse the ReplicaMap to find blocks on a specified volume if you want to remove a volume.

Reconfiguration of Hadoop

- HADOOP-7001 introduced a framework for reconfiguration of Hadoop
  - Triggered by Servlet
  - Reload specified configuration from config file
    - Keep the consistent between runtime configuration and the persistent configuration

Volume Refreshment

- Based on the framework provided by HADOOP-7001
- Implement a refresh method
  - Got the disks configured in files
  - Check which disk is in configuration file but not in service
  - Check and load the above volumes
  - Load blocks if there are some in the volumes

Future Works

- A disk rebalance function
  - Disk might be un-balanced, because:
    - Change failed disk
    - Mass delete some data
    - Consumed by other system, such as mapReduce local storage.
  - This function may relate to another…
- Monitor disk usage status of Datanode
  - Check data distribution among disks
  - Triggered by servlet
- Improve the functionality of invalidating a disk online and push it to mainline.
  - Challenges: when we configure the RAID card, it may conceal some details of disk details, which we observe by smartd. We need to do more verifications later.

Thank You!