Министерство науки и высшего образования Российской Федерации ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «Национальный исследовательский университет ИТМО» (Университет ИТМО)

ИТМО

Hadoop Distributed File System

Выполнила: Мигулаева Татьяна К4212с

Мигулаева Татьяна К4212с

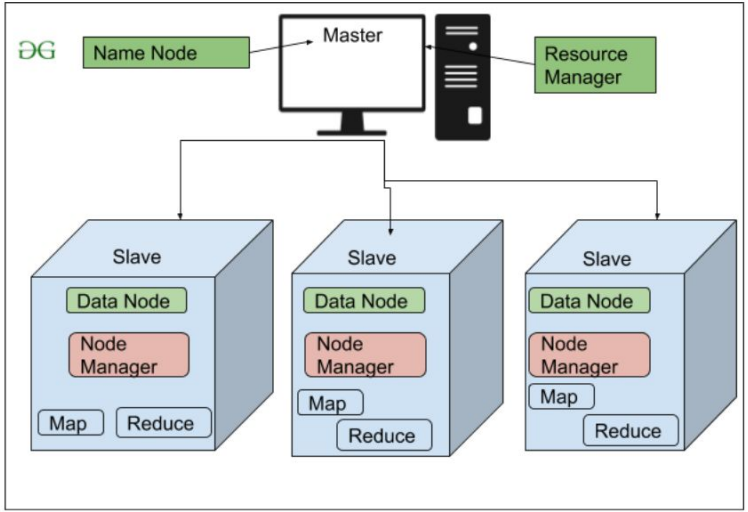
ИТМО

Актуальность

Увеличение объёмов больших данных в современных организациях приводит к серьёзным вызовам для традиционных систем хранения. Эти системы часто не способны эффективно справляться с требованиями высокой производительности, масштабируемости и гибкости, что критично для анализа больших данных. Традиционные хранилища также зависят от фиксированной инфраструктуры, что ограничивает их адаптацию к динамично меняющимся бизнес-требованиям.

Определение

HDFS (Hadoop Distributed File System) — это файловая система, предназначенная для работы с большими данными в экосистеме Hadoop на распределённых кластерах. HDFS разделяет данные на множество блоков и хранит их на серверах в кластере.



Мигулаева Татьяна К4212с

ИТМО

Принципы работы

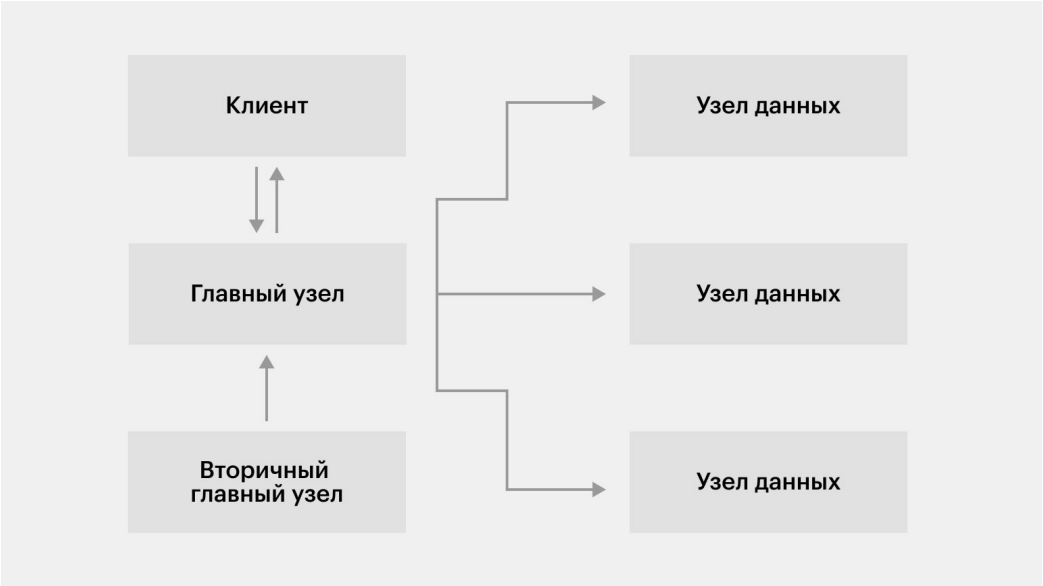
- HDFS разбивает файлы на небольшие блоки и хранит их на разных узлах в кластере серверов.
- Каждый блок данных в HDFS дублируется на несколько узлов.
- Обработка данных может идти в режиме реального времени в процессе их получения, что ускоряет работу
- HDFS легко масштабируется по горизонтали. Если объём данных или нагрузка увеличиваются, то можно просто добавить больше серверов в вычислительный кластер

Мигулаева Татьяна К4212с

ИТМО

Архитектура

- Клиент — это приложение, обеспечивающее взаимодействие пользователя через API с главным узлом, то есть с HDFS в целом.
- Главный узел управляет пространством имён файловой системы, хранит «карту» распределения файлов по блокам и их метаданные.
- Вторичный главный узел поддерживает актуальной собственную копию FSImage, периодически получая файлы EditLogs из NameNode.
- Узел данных - Это серверы, которые непосредственно работают с блоками данных



Преимущества SDS

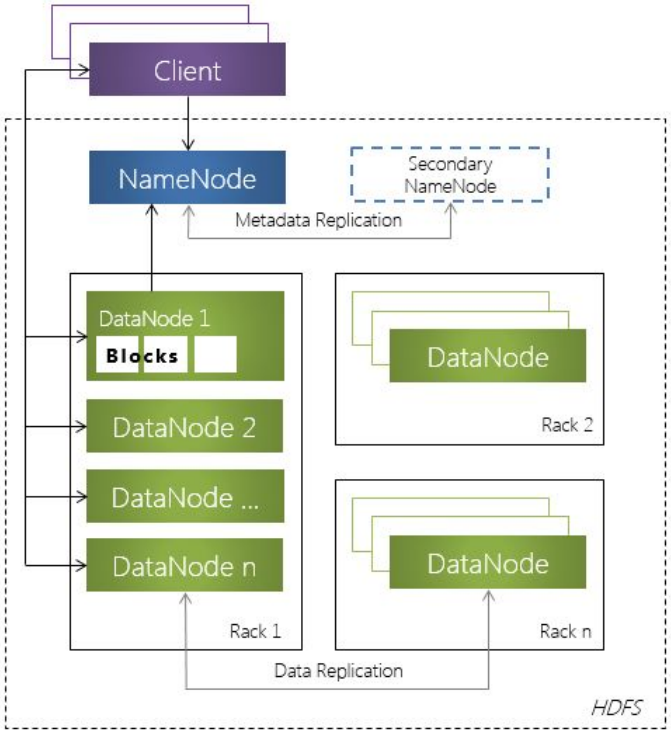
- Распределённое хранение
- Репликация данных
- Работа в формате потока данных
- Масштабируемость
- Поддержка различных типов данных
- Интеграция с экосистемой Hadoop

Недостатки и ограничения SDS

- Низкая эффективность работы с файлами меньше размера одного стандартного блока — 128 МБ
- Работа системы полностью зависит от главного узла. Если по какой-либо причине он перестанет работать, то вся HDFS выйдет из строя.
- Низкая безопасность данных

Пример реализации

Ответом GFS стал opensource проект Hadoop, включая Hadoop Distributed File System. Проект активно поддерживается и развивается компанией Yahoo. Facebook и Linkedin используют HDFS для хранения и анализа данных о пользователях, обеспечивая персонализированный опыт для каждого пользователя. Netflix использует HDFS в своей архитектуре Big Data для анализа просмотров и поведения пользователей, что помогает в принятии решений о создании нового контента. Amazon применяет HDFS как часть своих облачных решений на платформе Amazon EMR (Elastic MapReduce), что позволяет пользователям обрабатывать большие объёмы данных, используя инфраструктуру Amazon.



Мигулаева Татьяна К4212с

Заключение

HDFS (Hadoop Distributed File System) представляет собой мощное решение для хранения и обработки больших данных, обеспечивая высокую производительность, отказоустойчивость и масштабируемость. В условиях роста объёмов данных HDFS продолжает играть ключевую роль в развитии аналитики и обработки информации в современном бизнесе.

Спасибо за внимание!

Готова ответить на ваши вопросы

ITMO more than a UNIVERSITY

Выполнила: Мигулаева Татьяна К4212с