

№9 Дәріс. Үлестірілген файлдық жүйелер және деректерді сақтау жүйелері (GFS, HDFS)

Дәрістің мақсаты:

Үлестірілген файлдық жүйелердің (Distributed File Systems) негізгі принциптерін түсіндіру, Google File System (GFS) және Hadoop Distributed File System (HDFS) архитектурасын, жұмыс принципін, артықшылықтары мен кемшіліктерін қарастыру.

Кілт сөздер:

Distributed File System, GFS, HDFS, блок, репликация, NameNode, DataNode, fault tolerance, масштабталу, big data, master-slave architecture.

1. Кіріспе

Қазіргі ақпараттық жүйелерде үлкен деректер көлемі — Big Data негізгі факторға айналды. YouTube, Google Search, Facebook, Amazon сияқты жүйелер күн сайын **петабайттармен** (1 PB = 1000 TB) дерек өңдейді.

Осындай көлемдегі ақпаратты сақтау және өңдеу үшін үлестірілген файлдық жүйелер қолданылады.

2. Үлестірілген файлдық жүйе деген не?

Анықтама. Үлестірілген файлдық жүйе — бірнеше сервер дискілерінде жатқан файлдарды біртұтас логикалық жүйе ретінде көрсететін жүйе.

Негізгі мақсаттары:

- Үлкен көлемді деректерді сақтау
- Жоғары қолжетімділік пен сенімділік
- Масштабталу (кластерді ұлғайту)
- Ақауларға төзімділік

3. Үлестірілген файлдық жүйелердің негізгі қасиеттері

Қасиет	Сипаттамасы
Fault Tolerance	Сервер істен шықса да жүйе тоқтамайды
Scalability	Жаңадан сервер қосу арқылы жүйені үлкейту
High Throughput	Үлкен деректер ағынын өңдеу
Data Replication	Деректердің көшірмесін бірнеше жерге сақтау
Parallel Processing	Файлдарды параллель түрде оқу/жазу

4. GFS (Google File System)

GFS деген не?

GFS — Google жасаған үйреншікті файлдық жүйелерден әлдеқайда масштабты және ақауға төзімді жүйе. Google Search, Gmail, YouTube сияқты қызметтерде қолданылады.

Архитектурасы:

Компонент	Мәні
Master Server	Негізгі басқарушы түйін
Chunkservers	Дерек сақтайтын серверлер

Компонент	Мәні
Client	Файлмен жұмыс жасайтын қосымшалар

Маңызды түсініктер:

- Файл 64 MB+ блоктарға (chunk) бөлінеді
- Әр chunk 3 репликамен сақталады
- Деректер ChunkServer-де, ал метадерек Master-де

Артықшылықтары:

- Үлкен файлдарды тиімді өңдейді;
- Ақауларға төзімді;
- Жоғары өткізу қабілеті.

Кемшіліктері:

- Master-node — бір нүктелі ақау (single point of failure);
- Аз файлдар үшін тиімсіз.

5. HDFS (Hadoop Distributed File System)

HDFS деген не?

HDFS — Apache Hadoop шеңберінде жасалған, GFS идеясына негізделген ашық жүйе.

HDFS архитектурасы:

Компонент	Рөлі
NameNode	Метадеректерді басқарады
DataNode	Деректер блоктарын сақтайды
Secondary NameNode	Checkpoint жүргізеді (backup емес!)

Блок өлшемі. Әдетте 128 MB (кейде 64 MB немесе 256 MB)

Репликация коэффициенті.

Стандарт — 3.

Жұмыс принципі:

1. Файл блоктарға бөлінеді
2. Әр блок — 3 түрлі DataNode-ке жазылады
3. NameNode тек мета-ақпаратты сақтайды

6. GFS және HDFS салыстыру

Критерий	GFS	HDFS
Даму	Google (жабық)	Apache (ашық)
Архитектура	Master + ChunkServers	NameNode + DataNode
Файл өлшемі	Үлкен	Үлкен
Репликация	3	3
Негізгі қолдану	Google қызметтері	Big Data, Hadoop жүйелері
Flexibility	Аз	Жоғары

7. HDFS қолдану мысалдары.

Компания	Қолдану саласы
Yahoo	Web analytics
Facebook	Желі логтары, ML
LinkedIn	Kafka логтары
Netflix	Видео аналитика

8. Артықшылықтары мен шектеулері

Артықшылықтар:

- Үлкен файлдар үшін тиімді
- Масштабталатын архитектура
- Ақауларға төзімді
- Оқу жылдамдығы жоғары

Шектеулер:

- Кіші файлдар үшін тиімсіз
- NameNode — әлеуетті бір ақау нүктесі
- Моментальды (real-time) емес

Қорытынды

GFS және HDFS — үлкен деректерді сақтау үшін арналған ең танымал үлестірілген файлдық жүйелер. GFS — Google-дің жабық стандарты, ал HDFS — ашық көзі бар стандарт.

Бақылау сұрақтары

1. Үлестірілген файлдық жүйе деген не?
2. GFS архитектурасының негізгі компоненттері қандай?
3. HDFS-те NameNode қандай роль атқарады?
4. Репликация коэффициенті деген не?
5. GFS пен HDFS айырмашылығы қандай?
6. HDFS қандай салаларда қолданылады?
7. Неліктен үлкен блок өлшемі қолданылады?

Пайдаланылған әдебиеттер

1. Ghemawat, S., Gobioff, H., & Leung, S.-T. *The Google File System*, Google Research, 2003.
2. Shvachko, K. et al. *The Hadoop Distributed File System*, IEEE, 2010.
3. Tom White — *Hadoop: The Definitive Guide*, O'Reilly.
4. Tanenbaum, A. — *Distributed Systems: Principles and Paradigms*.
5. Apache Hadoop Documentation — hadoop.apache.org.
6. Jeff Dean lectures, Google.