Storage Systems

Cassandra, Bigtable, Dynamo and Google File System

Task 3 - Essay

Artem Sliusarenko, Mihailo Cvetkovic, Eugen Lucchiari Hartz

ID2221, Fall 2024, Period 1

October 26, 2024

Introduction

- **Challenges**: Managing large data volumes and high data throughput in distributed computing
- **Solutions:** Cassandra, Bigtable, Dynamo, and GFS
  - Improve scalability, fault tolerance, and performance
- **Impact:** Enable efficient data handling and high availability
  - Optimise data storage and retrieval for advanced analytics and real-time processing

Problem Description

- **Core Challenge**: Efficient management of vast data across distributed systems for high availability, performance, and scalability
- **Issues**: Traditional architectures falter with the required massive scale
- **Proposed Solution**: Architectural innovations for handling failures without data disruption
- **Our Focus**: Evaluation of solutions

Contribution

- **Cassandra**: Decentralised architecture, geographical spread
- **Bigtable**: Scalability, flexible data model
- **Dynamo**: Availability, efficient storage
- **GFS**: Fault tolerance, chunk storage

Solution Overview

- **Cassandra**: Dynamic data model, high throughput and reliability
- **Bigtable**: Distributed architecture, optimised for data handling
- **Dynamo**: Eventual consistency model, decentralised and scalable
- **GFS**: Built for scale and fault tolerance, efficient large-file management

Achievements

- **Cassandra**: Power extensive data management for Facebook, high efficiency in handling large-scale data environments
- **Bigtable**: Supports crucial applications at Google such as Maps and YouTube, flexibility and scalability in its architecture
- **Dynamo**: Ensures Amazon services remain uninterrupted, even during partitions
- **GFS** : Cost-effective base for significant Google data processes, foundation for systems like MapReduce

Comparative Evaluation - Strengths

- **Cassandra**: fault tolerance and scalability, excels in distributed applications
- **Bigtable**: fast data access and processing for services like Google Maps and YouTube
- **Dynamo**: Ensures continuous operation of Amazon's platforms, crucial for e-commerce resilience
- **GFS** : Cost-effective hardware use, ideal for large data processing tasks, supports systems like MapReduce

Comparative Evaluation - Weaknesses

- **Cassandra**: Prioritises availability over consistency, possibly unsuitable for applications that require strong consistency.
- **Bigtable**: Adaptability challenges outside Google-specific applications due to complexity
- **Dynamo**: Trades consistency for availability, problematic for applications requiring precise data correctness
- **GFS** : Optimised for large files, may underperform with small, frequent data needs of modern social media platforms

Conclusion

- **Impact**: Revolutionised data management in distributed systems
- **Solutions**: Unique approaches to performance, availability, and scalability
- **Influence**: Enhanced operations for major companies, set industry benchmarks

Questions?

Thank you for your attention!

ID2221, Fall 2024, Period 1

October 26, 2024