Verteilte Informationssysteme

Mitschrift von Aaron Winziers
Wintersemester 2019/20

1 Organisation

Prüfungstermine:

- 19.02.2020
- 01.04.2020

2 Introduction

2.1 Lifecycle of an information system

- Small company(startup) - Bookkeeping etc still possible on paper
- Growth - Information and data becomes too big for paper - database is needed
- Continued growth - Need for more databases or even specialized systems, more locations need more databases

2.2 What to do?

- Data warehouse
- Distributed architecture
- Replication - Which locations need which data?
- Cloud computing

There are 2 important use cases that should be considered: transactional load(OLTP)(write heavy) and analytical load(OLAP)(read heavy)

2.3 SQL vs NoSQL

2.3.1 SQL:

- (+) declarative queries
- (+) consistency
- (+) guarantees
- (+) data independence

- (+) normalization
- (-) do not scale well with increasing load
- (-) features not always needed/are not used
- (-) rigid data structures

2.3.2 NoSQL:

- ( ) not relational
- (+) scale well, are distributes in nature (more nodes = more performance)
- ( ) often w/o query language but with simple API
- (-) weak consistency models - distributed copies may not be identical
- (+) offer high performance

Both systems are typically combined into hybrid systems

2.4 Reasons for distributing data

- Cost and scalability - mainframes are difficult to extend
- Replications leads to higher availability
- Integration of different software modiles - prevent collisions
- Integration of legacy systems - old system can continue to exists parallel to new system
- New kinds of applications - esp. e-commerce
- Market forces

2.5 Why distribution?

Distributed data better corresponds to modern enterprise structures

3 Distributed query processing

3.1 Important aspects

More replications faster queries but slower updates

Fragmentation storing local data locally

Parallelism multiple queries can be performed at once, or a single query can be split into parts and executed at the same time

Transparency fragmentation, replication etc should not be need to be taken into account by the user

3.2 Systems differ in terms of:

- Degree of coupling
- Interconnection structure
- Interdependence of components
- Synchronization of components

3.3 Forms of distributed systems

- Peer-to-Peer and file sharing
- Cloud computing
- Web services and the deep Web
- Semantic Web
- Big Data Analytics

3.4 Fallacies of distributed computing:

- The network is reliable
- Latency is zero
- Bandwidth is infinite
- The network is secure
- Topology doesn't change
- There is one administrator
- Transport cost is zero
- The network is homogenous
- Location is irrelevant

4 Fragmentation and allocation in distributed database management systems

5 Replication and synchronization

6 Grid and cloud computing

7 Distributed transactions

8 Information integration

9 Distributed information retrieval

10 Parallel database systems