Cloud Data Management

Seminar WS 2009/10

Seminar: Anrechnungsmöglichkeiten

- Masterstudium
  - Teil der Kern- bzw. Vertiefungsmodule *Moderne Datenbanktechnologien (*bzw. *Anwendungsspezifische Datenbankkonzepte)*
  - *Seminarmodul* oder *Masterseminar*
- Bachelorstudium
  - *Seminarmodul* oder *Bachelorseminar*
- Alte Studiengänge (Diplom, etc.)
  - Problemseminar

Cloud Coumputing

"Cloud computing is using the internet to access someone else's software running on someone else's hardware in someone else's data center"

- Lewis Cunningham

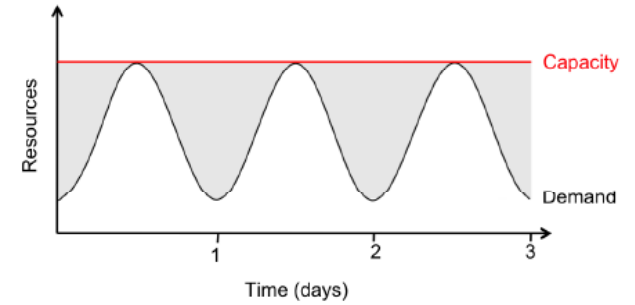
- Externe Bereitstellung von IT-Infrastrukturen sowie Applikations-Hosting über das Internet (bzw. Intranet)
- Public Cloud vs. Private Clouds

Cloud Computing (2)

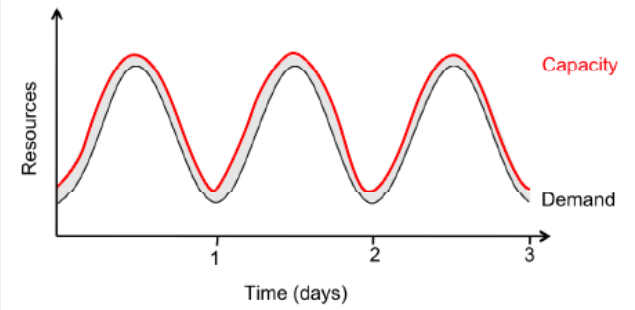
- Charakteristika:
  - Illusion unendlicher, „on demand" verfügbarer Ressourcen
  - Virtualisierung - gemeinsame Nutzung v. Ressourcen durch viele Nutzer
  - „Elastizität" - schnelle Belegung/Freigabe von Ressourcen nach Bedarf („Hinzuschalten" weiterer Rechner)
  - Abrechnung nach Verbrauch (CPU Zyklen, Speicherplatz, Übertragungsvolumen)

Cloud Computing – Cloud User

- kein Einrichten/Betreiben eigener Rechenzentren
- keine langfristige Ressourcenplanung
- keine hohen Vorabinvestitionen - pay per use
- Verspricht wesentliche Kosteneinsparungen



eigenes Rechenzentrum



Rechenzentrum in der Cloud

Bildquelle: M. Armbrust, et al.: .Above the clouds: A Berkeley view of cloud computing. Technical report, February 2009.

Cloud Computing – Cloud Provider

- Aufteilung verfügbarer Ressourcen auf mehrere Kunden
- große Rechenzentren (50.000 Server) haben im Vrgl. zu mittelgroßen (1000 Server) nur 1/5 - 1/7 der Kosten
- Standortvorteile - Elektrizitätspreise, Löhne, Steuern
- Green Computing
  - bessere Auslastung der Clouds im gegenüber lokalen Rechenzentren

Cloud Computing - Einteilung

SaaS
Software as a Service

PaaS
Platform as a Service

IaaS
Infrastructure as a Service

Infrastructure as a Service (IaaS)

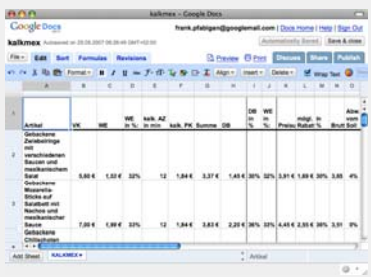
- Bereitstellung von IT-Infrastruktur
  - Rechner
  - Betriebssysteme
  - Firewalls, Router, …
- *Amazon EC2 (Electronic Compute Cloud)*- „Mieten“ einer Menge virtueller Maschinen, auf denen Anwendungen jeder Art zur Ausführung gebracht werden können

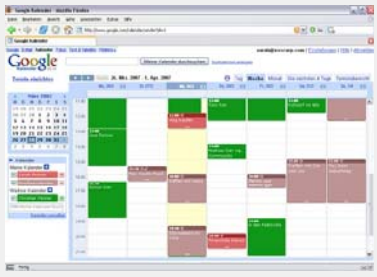
Platform as a Service (PaaS)

- Bereitstellung von Entwicklungs- und Hosting-Umgebung
- *Google App Engine* - Plattform zum Entwickeln und Vorhalten von Webanwendungen
- *Amazon S3* - Speichern beliebig großer Datenmengen

Software as a (Cloud) Service

- Beispiele: Google/Yahoo Mail, Google Docs, Google Calendar, Salesforce.com
- Anwendungen werden als Dienste über das Internet bereitgestellt
- keine lokale Software-Installation, Datenhaltung
- universeller Zugang mittels Browser/Thin Client





Cloud Computing - Einsatzfälle

- E-Commerce Startups
  - keine eigenen Ressourcen einzurichten
- „Online-Speicher“
  - Amazon S3 - Speichern und Lesen von Daten über einen Webservice von überall zu jeder Zeit
  - schnell, billig, hochverfügbar, hochskalierbar
- Paralleles OLAP
  - multidim. Aggregation großer Datenmengen
  - Data Mining
  - ACID nicht erforderlich

Cloud Data Management

- Cloud-Anwendungen erfordern Datenhaltung / Austausch von Daten in der Cloud
- effiziente und persistente Verwaltung großer Datenmengen (mehrere TB bis PB)
- Möglichkeiten:
  - Filesystem (Google File System, Hadoop DFS)
  - Cloud Datenbanken
    - Key value stores (Bigtable, SimpleDB, …
    - relationale Datenbanken
- Map/Reduce-Paradigma zur Parallelisierung

Google File System (GFS)

- verteiltes, hochskalierendes, Linux-basiertes Dateisystem (tausende Festplatten, mehrere 100TB)
- optimiert für Google Search Engine
- File-Änderungen durch Anhängen
  - kaum Änderungen innerhalb eines Files
- Daten werden in 64 MB Chunks repliziert gespeichert (mind. 3 Replikate)
- Master-Server für Chunk-Verteilung und Zugriffsverwaltung

Google Bigtable

- verteilte Verwaltung strukturierter Daten
- hochskalierend (mehrere PB), verwendet GFS
- Gmail, Google Maps, Google Earth, Web Indexe verwenden Bigtable
- Key-Value-Store, keine voll relationales Modell
  - Eingeschränkte Anfragen (GQL)
  - read/write nur für eine Zeile atomar

MapReduce

- Google Framework zur automatischen Parallel-isierung von Auswertungen auf großen Datenmengen
- OpenSource-Implementierung Hadoop
- Nutzung v.a. zur Verarbeitung riesiger Mengen teilstrukturierter Daten (Google,Yahoo,Facebook…)
  - *Konstruktion Suchmaschinenindex*
  - *Clusterung von News-Artikeln*
  - *Spam-Erkennung …*
- Datenparallelisierung über zwei Funktionen zur Partitionierung (Map) und Kombination (Reduce)

MapReduce (2)

- Map-Funktion:

$$(K_{in}, V_{in}) \rightarrow \text{list}(K_{inter}, V_{inter})$$

- Reduce-Funktion:

$$(K_{inter}, \text{list}(V_{inter})) \rightarrow \text{list}(K_{out}, V_{out})$$

- Beispiel: paralleles Zählen der Vorkommen aller in einer Dokumentmenge enthaltenen Wörter

```
map(String key, String value):
  // key: document name
  // value: document contents
  for each word w in value:
    EmitIntermediate(w, "1");
```

```
reduce(String key, Iterator values):
  // key: a word
  // values: a list of counts
  int result = 0;
  for each v in values:
    result += ParseInt(v);
  Emit(AsString(result));
```

Quelle: J. Dean and S. Ghemawat. MapReduce: Simplified data processing on large clusters.Proc. OSDI, 2004, pages 137–150

- MapReduce und Datenbanken ?

SEMINAR

Seminarziele

- Beschäftigung mit einem praxis- und wissenschaftlich relevanten Thema
  - kann Grundlage für Abschlussarbeit oder SHK-Tätigkeit sein
- Erarbeitung + Durchführung eines Vortrags unter Verwendung wissenschaftlicher (englischer) Literatur
- Diskussion
- Schriftliche Ausarbeitung zum Thema
- Hilfe und Feedback durch zugeteilten Betreuer

Scheinvergabe / Modulprüfung

- selbstständiger Vortrag mit Diskussion (45 Minuten)
  - Abnahme der Folien durch Betreuer
- schriftliche Ausarbeitung (ca. 15 Seiten)
  - Abnahme der Ausarbeitung durch Betreuer
  - Ausarbeitung soll zum Vortragstermin vorliegen (Vorträge ab Januar 2010)
- aktive Teilnahme an allen Vortragsterminen

Seminar (3)

- Max. 20 Teilnehmer
- Vortragstermine
  - Montags, 11:00–12:45 Uhr, JG 1-22, ab 4. 1. 2010
  - Dienstags, 9:00–10:45 Uhr, FK-Hörsaal, ab 5. 1. 2010
- Themenzuordnung
  - Koordinierungstreffen mit Betreuer in den nächsten 2 Wochen, d.h. bis **spätestens 02.11.09**
  - ansonsten verfällt Seminaranmeldung
  - freiwilliger Rücktritt auch bis max. 2.11.09 (danach wird das Seminar als nicht bestanden gewertet)

Themen

Nr.	Thema	Termin	Betreuer	Studenten
1	Einführung Cloud Computing	4.1.	Sosna	
2	Cloud-Infrastrukturen (Amazon EC2 etc.)	4.1.	Thor/Aumüller	
3	**Cloud-Software-Plattformen (Salesforce, Google App Engine, Windows Azure)**	5.1.	Thor/Aumüller	
4	Sicherheit in der Cloud	11.1	Sosna	
5	Verteilte Dateisysteme in der Cloud (Google FS, Hadoop FS, Amazon S3)	11.1	Endrullis	
6	Database as a Service: Übersicht	12.1.	Maßmann	
7	ACID Transaktionen in Cloud-Umgebungen	12.1.	Endrullis	
8	**Key-Value-Stores (CouchDB, Bigtable, SimpleDB, Amazon Dynamo, Facebook Cassandra)**	18.1.	Groß	
9	Relationale Cloud-DB (SQL Azure, DB2 on EC2)	19.1	Kolb	
10	Multi-Tenant-Datenbanken für SaaS (Schema-Management)	19.1.	Maßmann	
11	**MapReduce : Konzept (Google, Hadoop)**	25.1.	Kolb	
12	Parallele Datenanalyse/-verarbeitung mit M/R (Pig!, SystemT/JAQL, Hive)	26.1	Köpcke	
13	Automatische M/R-Nutzung (HadoopDB, PigLatin)	26.1.	Köpcke	
14	**DB/UDF und M/R (**SQL/MapReduce, CouchDB)	1.2.	Hartung	
15	Parallele Cloud-DBS (Bigtable,SimpleDB, PNUTS)	2.2.	Groß	
16	Servicebasierte Datenintegration	2.2.	Hartung	