Dr Greg Wadley

THE UNIVERSITY OF MELBOURNE

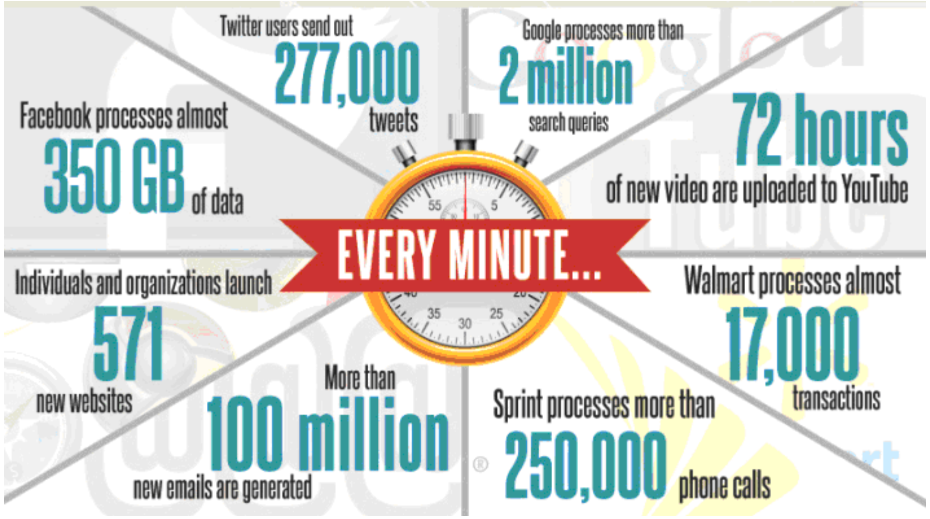
INFO90002 Database Systems & Information Modelling

Week 12

Database Trends

Today's Session…

- Relational status-quo
  - Major vendors
  - Current offerings
  - Market share
- Challenges to the relational status quo
  - Big Data
  - Cloud storage
  - Object-oriented applications
- Responses
  - NoSQL
  - NewSQL
  - Hadoop
  - In-memory databases



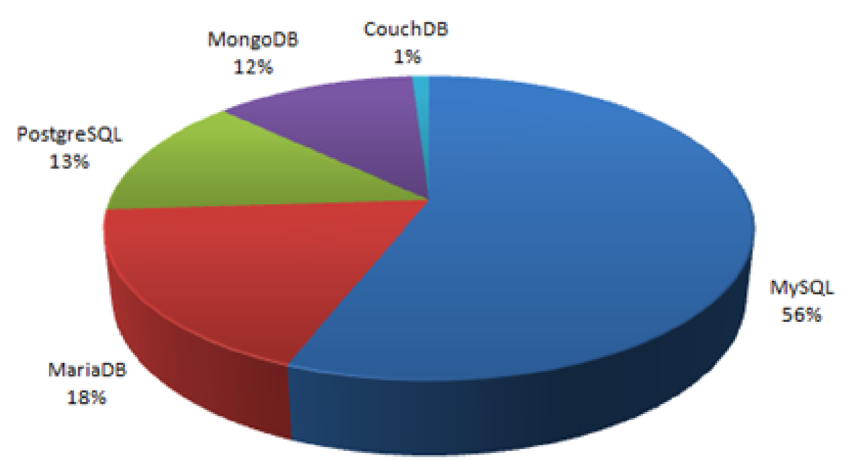
The Relational status-quo

- major vendors
- current offerings
- market share

Leading DBMS vendors

- Top *commercial* DBMS (Gartner, 2015)
  - 41.6% Oracle
  - 19.4% Microsoft Sql Server
  - 16.5% IBM DB2
  - (top 3 own over ¾ of market)
  - (top 5 NoSQL vendors together are at 8th place)

Open source DBMS

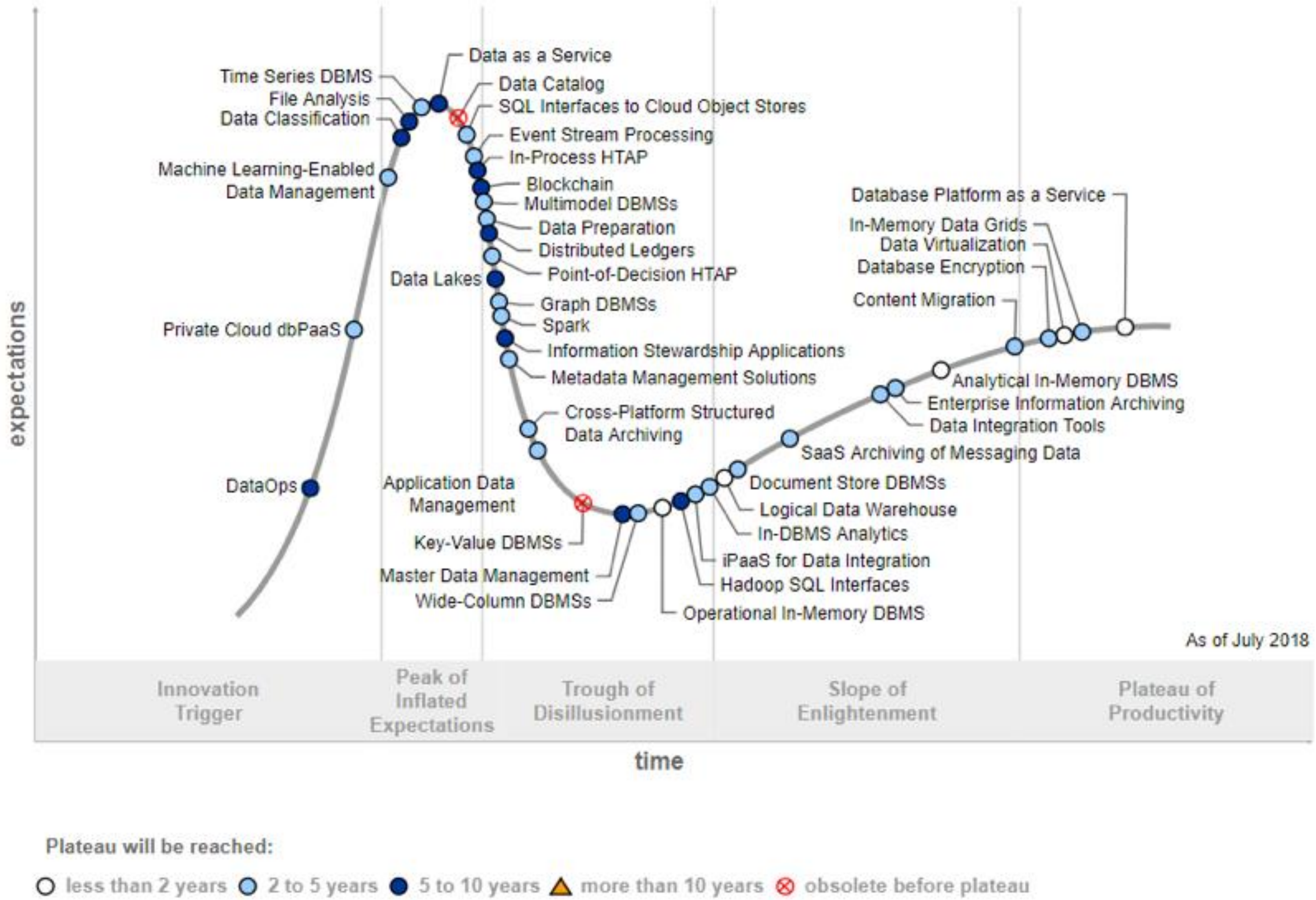


http://blogs.gartner.com/merv-adrian/2016/04/12/dbms-2015-numbers-paint-a-picture-of-slow-but-steady-change/

DBMS popularity

- **Number of mentions of the system on websites**, measured as number of results in search engines queries. At the moment, we use Google and Bing for this measurement. In order to count only relevant results, we are searching for <system name> together with the term database, e.g. "Oracle" and "database".
- **General interest in the system.** For this measurement, we use the frequency of searches in Google Trends.
- **Frequency of technical discussions about the system.** We use the number of related questions and the number of interested users on the well-known IT-related Q&A sites Stack Overflow and DBA Stack Exchange.
- **Number of job offers, in which the system is mentioned.** We use the number of offers on the leading job search engines Indeed and Simply Hired.
- **Number of profiles in professional networks, in which the system is mentioned.** We use the internationally most popular professional network LinkedIn.
- **Relevance in social networks.** We count the number of Twitter tweets, in which the system is mentioned.

Gartner's "Hype Cycle"



https://www.gartner.com/en/newsroom/press-releases/2018-09-11-gartner-hype-cycle-for-data-management-positions-three-technologies-in-the-innovation-trigger-phase-in-2018

Gartner's "Magic Quadrant"

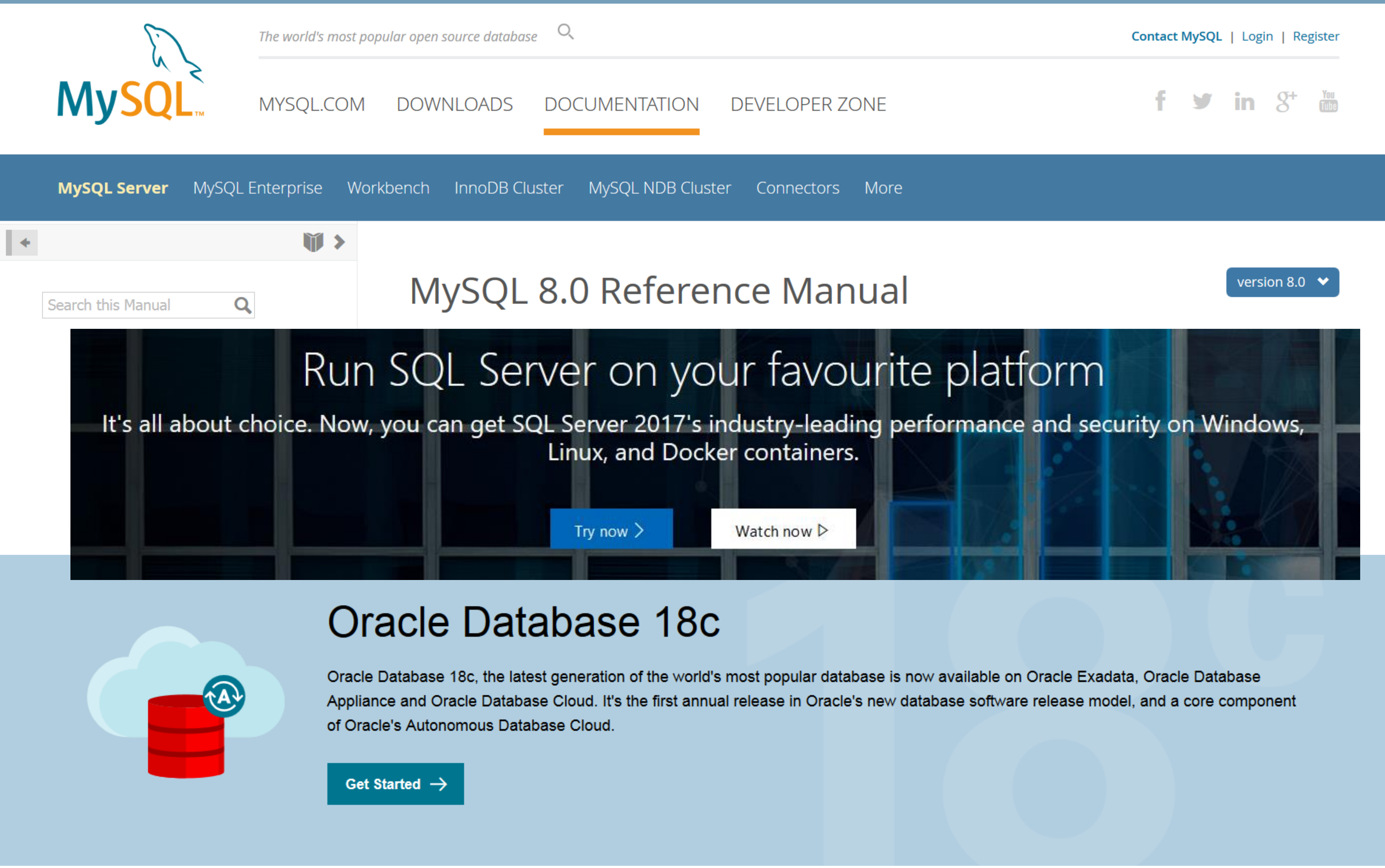
"In this Magic Quadrant, Gartner evaluates the strengths and weaknesses of 11 providers that it considers most significant in the marketplace, and provides readers with a graph (the Magic Quadrant) plotting the vendors based on their ability to execute and their completeness of vision. The graph is divided into four quadrants: niche players, challengers, visionaries, and leaders. … Gartner drastically changed revenue requirements for inclusion, and as a result, axed 10 providers. … Microsoft has put a little distance between itself and Oracle for the top spot in the space. ... SAP took a minor step back from the pack on the vertical axis, allowing Amazon Web Services to sneak through and supplant itself as the only real competition Microsoft and Oracle have left."

(Solutions Review, 2017)

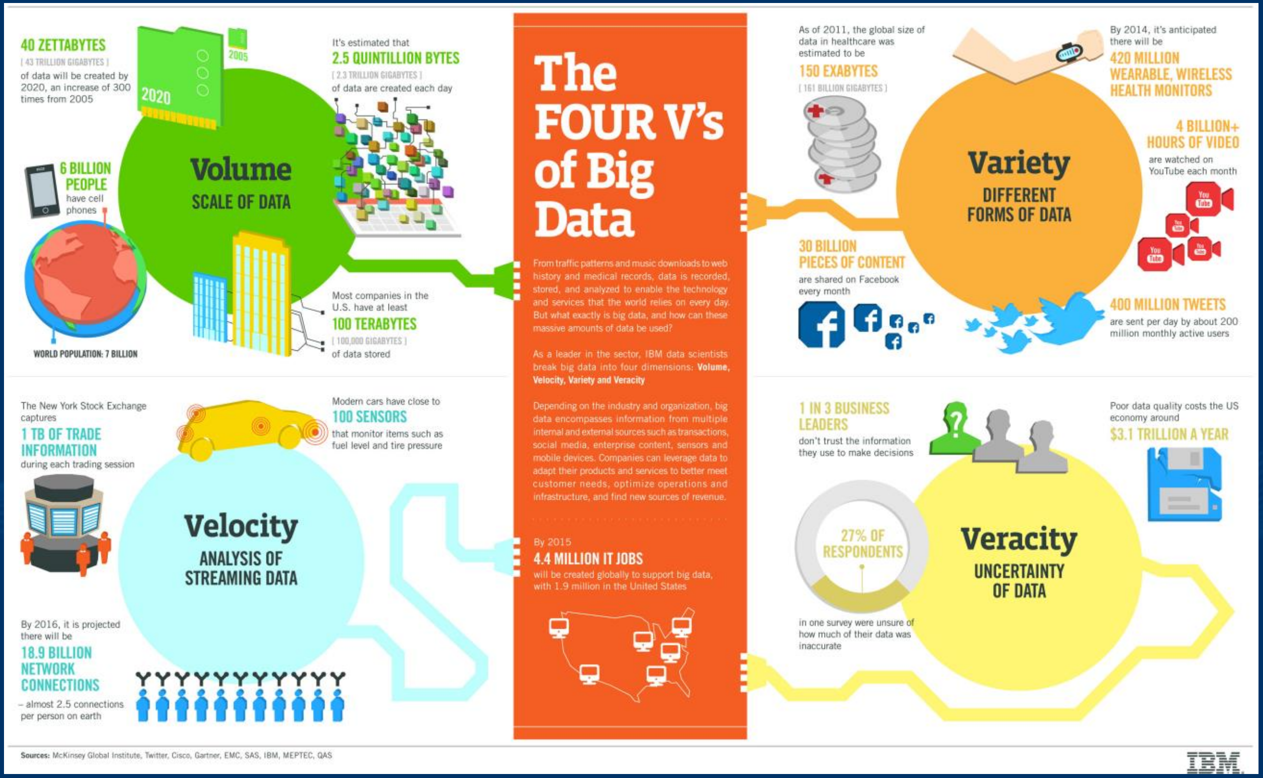
Source: Gartner (November 2017)

https://solutionsreview.com/data-management/whats-changed-2017-gartner-magic-quadrant-for-operational-database-management-systems/

Current major relational offerings



Challenge: Big Data



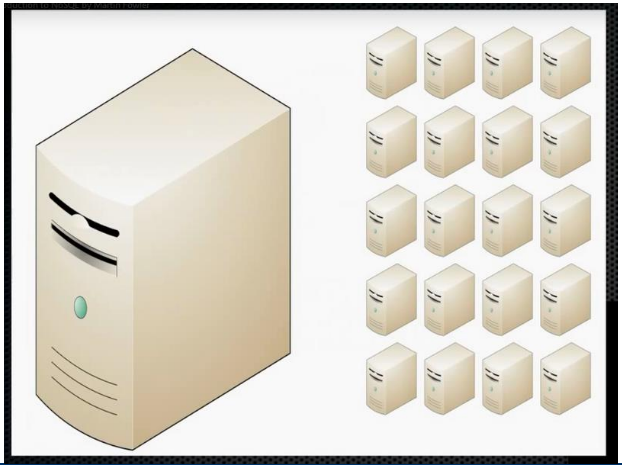
Extremely large data sets

- Difficult to process high volume and velocity of data using traditional database management and processing tools
- Challenges include capture, storage, search, sharing, analysis, and visualization.
- Relational databases emerged when data was generated within organizations, mostly manually by employees. This scenario still occurs, and relational may be best here.
- But data is now generated by: social media, global website clickstreams, machinery, sensors, scientific instruments.
- “Internet of Things” … 30 billion wireless devices by 2030 (http://www.dbta.com/Editorial/Trends-and-Applications/Powering-the-Internet-of-Things-with-Real-Time-Hadoop-103469.aspx)
- Exabytes … 1,000,000 terabytes 1,000,000,000 gigabytes
- Volume and velocity of data is too high for SQL databases

Solution: distributed database

- Large internet companies like Google and Amazon found they could not store huge data sets in relational databases
- How to scale up capacity?
- Vertical scaling (bigger servers) gives diminishing returns
- Therefore horizontal scaling:

  spread data across many small servers
- But since relational is not so good with distributed data, these companies devised new databases that are designed from the ground up to run across multiple servers
- Google: *Bigtable*
- Amazon: *Dynamo*
- inspired the NoSQL movement



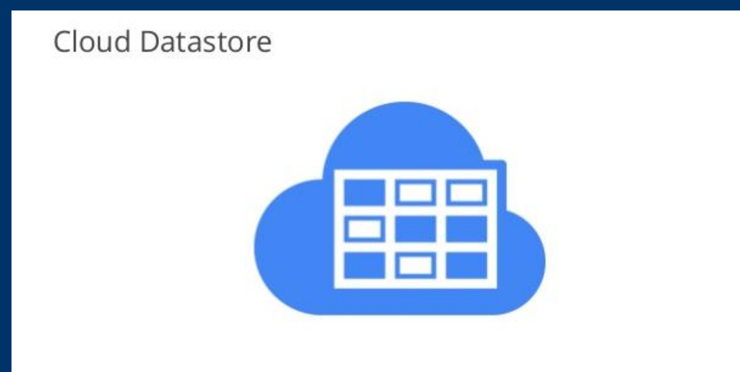
Big Data – problems

- In its infancy
  - tools for using and analysing big data, as well as standards, are still being developed
- Need the right talent and technology and structure of workflows to optimize the use of big data
- Requires expensive professionals ("data scientists")

  "There will be a shortage of talent necessary for organizations to take advantage of big data. By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions" (McKinsey Global Institute)

- https://mbs.edu/education-development/degreeprograms/masterofbusinessanalytics
- http://science-courses.unimelb.edu.au/study/degrees/master-of-data-science/

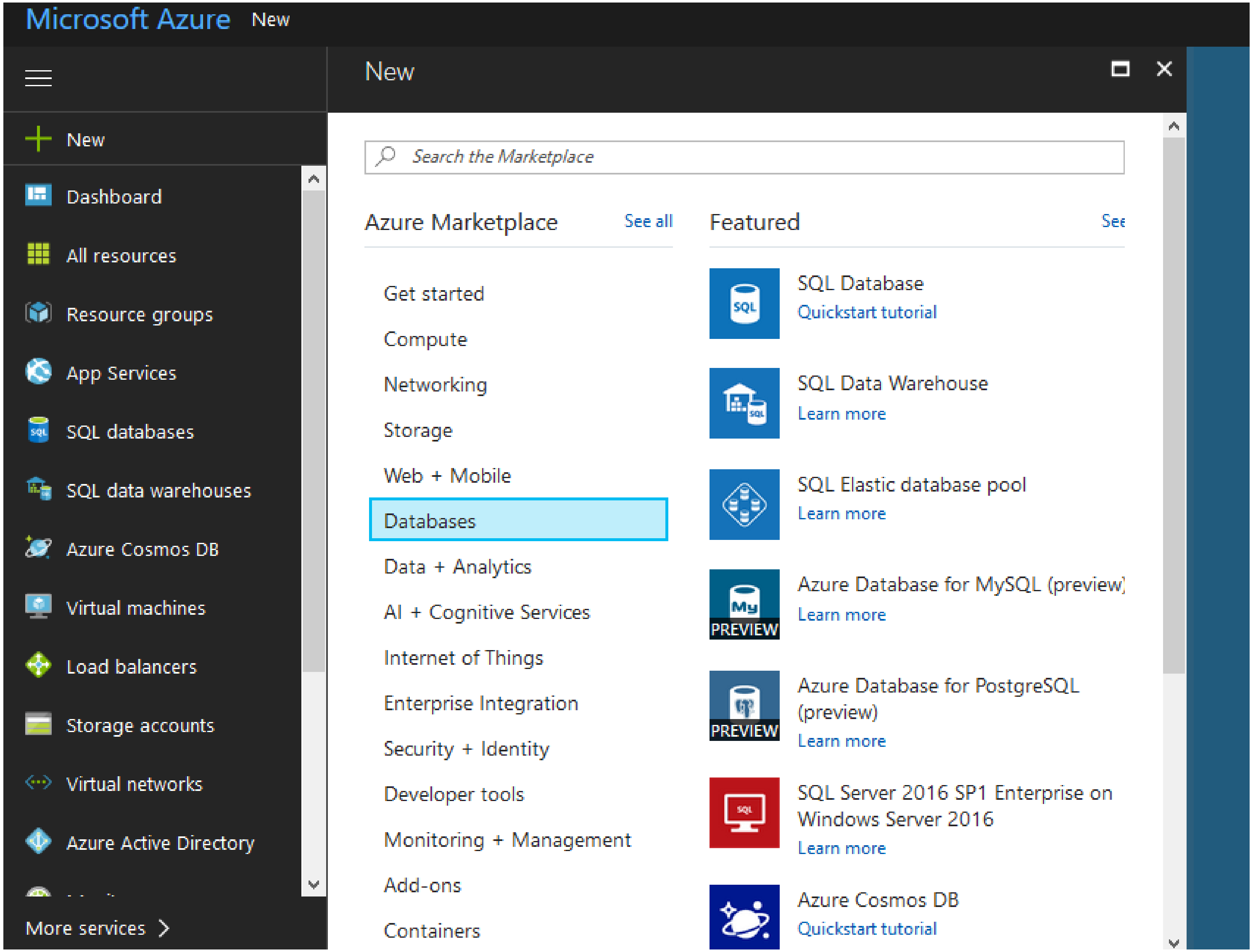
Challenge: Cloud storage



Database as a service (DBaaS)

- Similar to other cloud services
  - DBMS and data are in the cloud
  - your application connects as required
  - pay per usage: quantity data stored, input/output
  - DBMS is administered by cloud provider
    - reduces need for in-house DBA
    - may be managed via web console
- Can be presented as either
  - database as a service (DBaaS)
    - relational database
    - non-relational (NoSQL) database
  - Virtual machine with a database installed
- Advantages
  - Simplifies setting up, and especially scaling up, your database

Example: Microsoft Azure



Potential cloud db issues

- Data security
  - provider may not have fully integrated security structure
  - may need to resort to encryption
- Legal frameworks
  - need to (continue to) conform to laws governing use of data
    - Can the data be moved out of a political jurisdiction?
    - How must the data be secured?
- Large movements of data between your site and cloud
  - during initial setup
  - during ongoing integration with local data

sourced from InfoWorld (2012)
http://www.infoworld.com/d/cloud-computing/the-unpleasant-truths-about-database-service-208450

Example cloud database vendors

- Information from www.mitre.org/work/tech_papers/2012/11_4727/cloud_database_service_dbaas.pdf
- SQL
  - Amazon Relational Database Service
  - Microsoft SQL Azure
  - Oracle Cloud
  - Rackspace
- NoSQL
  - Google Datastore
  - Amazon SimpleDB
  - Microsoft Azure DocumentDB

Comparison of cloud db

	Amazon RDS (MySQL)	Microsoft SQL Azure	Google Datastore	Amazon SimpleDB
Type	**RDBMS**	**RDBMS**	**NoSQL**	**NoSQL**
Maximum amount of data that can be stored	1 terabyte per database[2]	50 gigabytes per database[3]	Not published for entire database, but 1 MB limit on a subset of data (called an entity). Limit to the number of indexes.	10 gigabytes per database domain (roughly equivalent to an RDBMS table)[4]
Ease of software portability with similar, locally hosted capability	High. MySQL instantiation in cloud is very similar to the local instantiated version.	High. Most SQL Server features are available in SQL Azure.	Medium/Low. Requires Java Data Objects or Datastore-specific interface and use of App Engine.	Medium. Requires SimpleDB-specific interface.
Transaction capabilities	Yes	Yes	Yes	Yes
Configurability and ability to tune database	High. MySQL instantiation in cloud.	Medium. Can create indexes and stored procedures, but no control over memory allocation or similar resources.	Low	Low
Database accessible as "stand-alone" offering	Yes	Yes	No. Requires Google App Engine application layer.	Yes
FISMA Certified	No	No	No	No
Can designate where data is stored (e.g., region or data center)	Yes	Yes	No	Yes
Replication	Yes	Yes	Yes	Yes

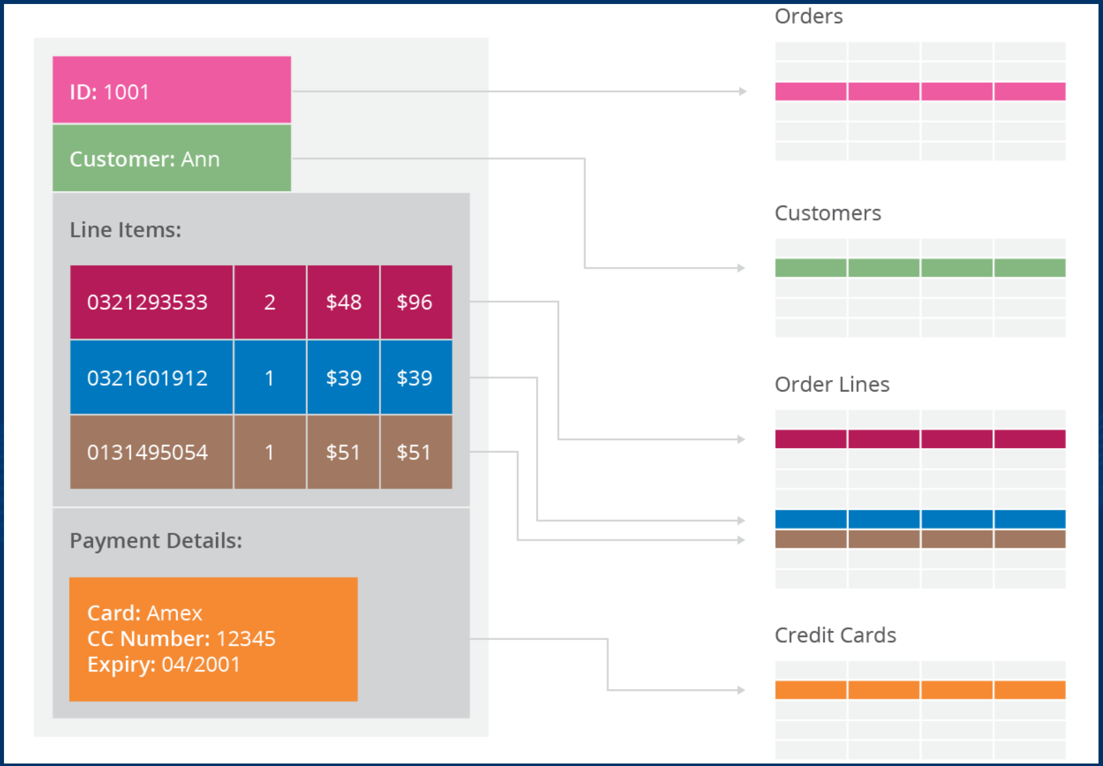
Pricing & Performance

	Amazon RDS (MySQL)	Microsoft SQL Azure	Google Datastore	Amazon SimpleDB
Example Pricing for Processing (Refer to Sections 2-5 for details)*	Ranges from $0.11 per RDS hour for smallest instance to $2.60 per hour for largest instance	Ranges from $9.99 per database with up to 1 GB of storage to $499.95 per database with up to 50 GB of storage per month	$0.10 per App Engine CPU hour (required for accessing Datastore)	$0.14 per SimpleDB unit hour
Example "On-demand" Pricing for Data Transfers (Refer to Sections 2-5 for details)*	Inbound $0.10 per GB and outbound ranges from $0.15 per GB to $0.08 per GB, depending on volume	Inbound $0.10 per GB Outbound $0.15 per GB	Inbound $0.10 per GB Outbound $0.12 per GB	Inbound $0.10 per GB and outbound from $0.15 to $0.08 per GB, depending on volume.
Example Monthly Pricing for data storage (Refer to Sections 2-5 for details)*	$0.10 per GB plus $0.10 per 1 million I/O requests	Included in processing pricing	$0.15 per GB	$0.25 per GB

Single Client	Average writes per second	Average reads per second
Amazon MySQL RDS*	2,567	2,551
Microsoft SQL Azure	406	410
Google Datastore**	288	200
Amazon SimpleDB	208	63

5 Clients	Average writes per second	Average reads per second
Amazon MySQL RDS*	7576	7905
Microsoft SQL Azure	1737	1893
Google Datastore**	N/A	N/A
Amazon SimpleDB	689	281

Challenge: Object-orientated application software



Relational vs OO mismatch

- Relational databases
  - Individual entities are represented by a simple row structure
- Object-oriented programming
  - OO allows more complex data structures
  - complex objects must be normalized to fit relational tables
- How can an OO program persist data in a database?
  - relational database (simple, powerful, but not OO)
  - object-relational databases (complex data types, SQL 1999)
  - object-oriented databases (did not become popular)
  - object-relational mapping software (currently popular)
  - NoSQL (becoming popular)

Object-oriented DBMS

- Direct storage of object-oriented data
- Offers benefits of OO and Relational
  – directly persists objects
  – offers performance and security of relational DB
- Products include:
  – Db4o, InterSystems, McObject, Objectivity, Progress, Versant, ObjectDB, Gemstone, VelocityDB …
- But did not become popular in the marketplace
  – often tied to particular OO programming languages
  – lack of compatibility between different OODMBSs
  – lack of standard ad hoc query language
  – RDBMS are used to integrate applications
  – (but, re-emerging in the form of NoSQL ‘document databases’)

Responses

“The database market is back in play after a 30-year old freeze in which Oracle dominated the high end, and Microsoft the mid market. Then along came open source, the cloud, NoSQL, in memory and everything changed….The idea that everything is relational? Those days are gone.”

Redmonk analyst James Governor

NoSQL

- Features
  - doesn’t use relational model or SQL language
  - designed to run on distributed servers
  - most are open-source
  - built for the modern web
  - schema-less (though there may be an "implicit schema")
  - 'eventually consistent‘
  - tend not to offer relational features like ACID, locks, joins
- Goals
  - to improve programmer productivity (OR mismatch)
  - to handle larger data volumes and throughput (big data)

from *NoSQL Databases: An Overview*
by Pramod Sadalage, Thoughtworks (2014)

NewSQL

- Features
  - like NoSQL, but offers ACID transactions
  - suited where consistency is important, e.g. financial data
  - offers traditional relational features like SQL, transactions
  - but scales horizontally like NoSQL systems
  - distributed, automatic sharding

- Example products
  - H-Store
  - VoltDB
  - Spanner

	CLOUD SPANNER	TRADITIONAL RELATIONAL	TRADITIONAL NON-RELATIONAL
Schema	✓ **Yes**	✓ Yes	✗ No
SQL	✓ **Yes**	✓ Yes	✗ No
Consistency	✓ **Strong**	✓ Strong	✗ Eventual
Availability	✓ **High**	✗ Failover	✓ High
Scalability	✓ **Horizontal**	✗ Vertical	✓ Horizontal
Replication	✓ **Automatic**	Configurable	Configurable

In-memory databases

- data are stored in memory and processed there
- disk is for long-term persistence
- primary motive is ***speed*** – of storage, transactions, analytics
- access speed of main memory is much faster than disk
- SAP HANA is the best-known example
- can use up to 100 terabytes memory
- supports OLTP and OLAP
- appeared in 2010
- (there are other in-memory databases)

Hadoop

- for data-processing, not operational database
- Hadoop = framework of software utilities including:
  - HDFS: (Hadoop Distributed File System) for distributed processing across a cluster of many computers
  - Hbase: database for data storage
  - MapReduce: for processing data across the cluster

- based originally on Google technology
- about 10 years old
- used by Yahoo, Facebook, available on cloud