RDBMS and Big Data

Database Technologies and trade-offs

Nathan Cunningham

Associate Director for Big Data Network Support

University of Essex

UK Data Service

Outline

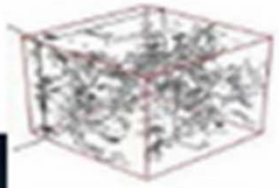
- Data Intensive Research – Complex data
- Preserving Data Collections
- Data and information products
- Trade offs
- Data as a service – a unified information architecture
- Fundamental preservation challenges for complex data types

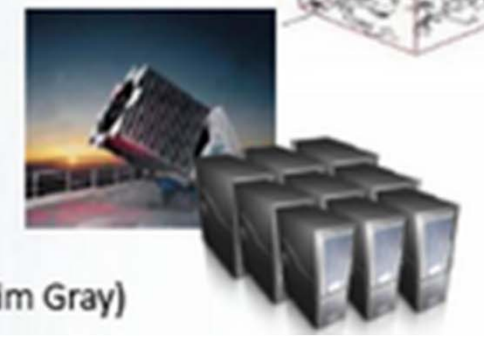
Data Intensive Research

Emergence of a Fourth Research Paradigm

1. Thousand years ago – **Experimental Science**
   - Description of natural phenomena
2. Last few hundred years – **Theoretical Science**
   - Newton's Laws, Maxwell's Equations…
3. Last few decades – **Computational Science**
   - Simulation of complex phenomena
4. Today – **Data-Intensive Science**
   - Scientists overwhelmed with data sets from many different sources
     - Data captured by instruments
     - Data generated by simulations
     - Data generated by sensor networks
   - ➢ eScience is the set of tools and technologies to support data federation and collaboration
     - For analysis and data mining
     - For data visualization and exploration
     - For scholarly communication and dissemination

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$

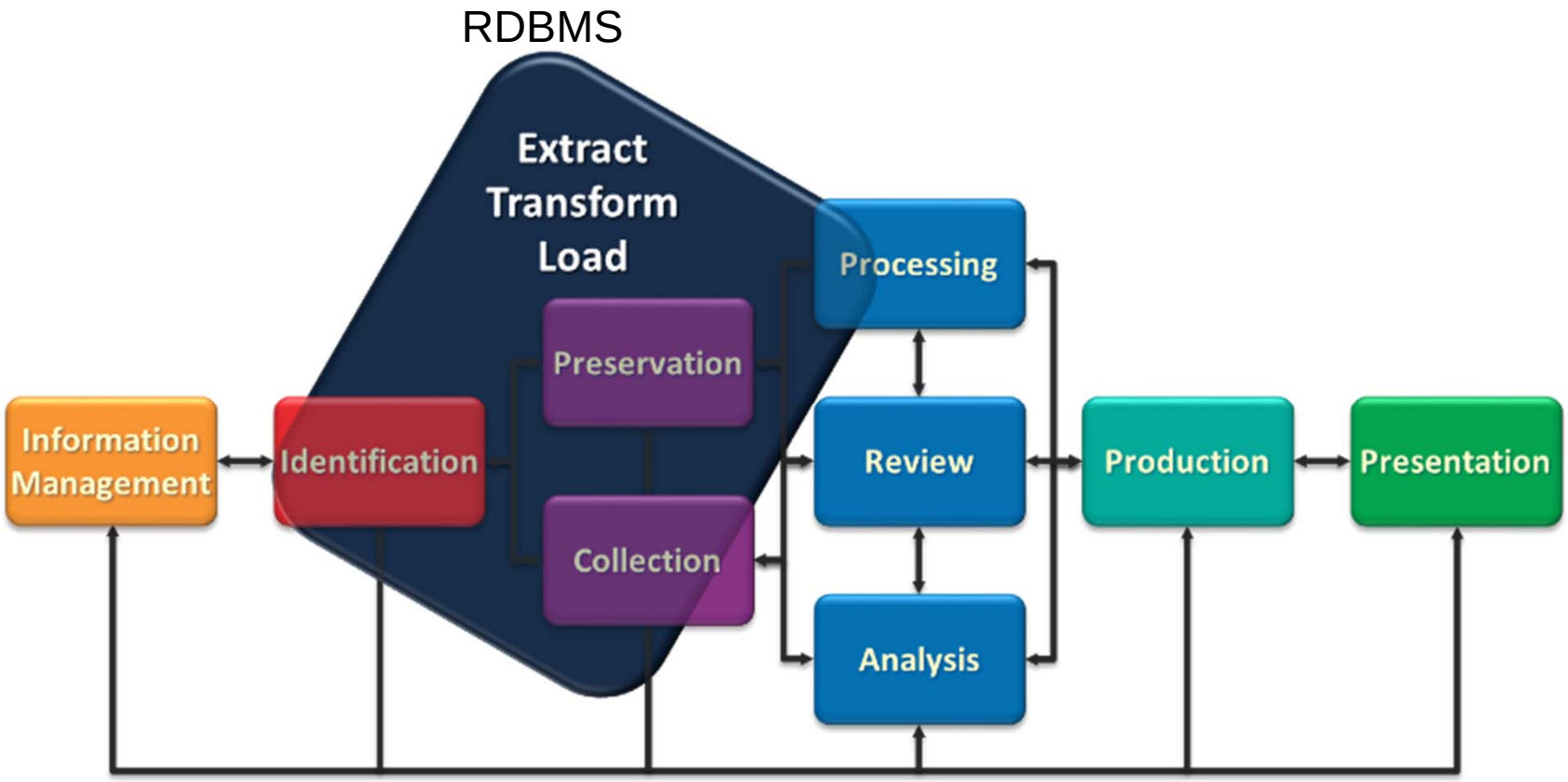




(With thanks to Jim Gray)

(2007) The Forth Paradigm, Tony Hey, Microsoft Research Labs

The Preservation Model

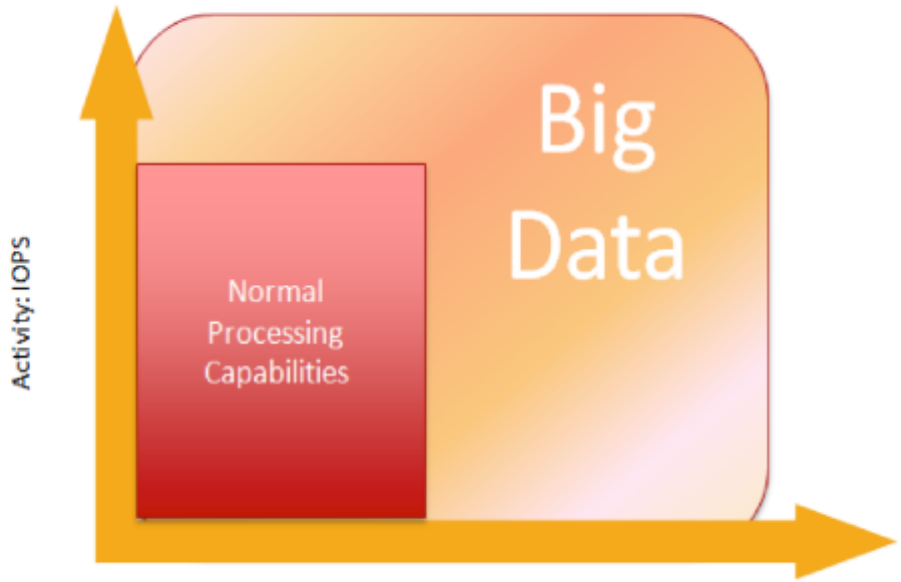


The Sedona Conference® Database Principles Addressing the Preservation and Production of Databases and Database Information in Civil Litigation

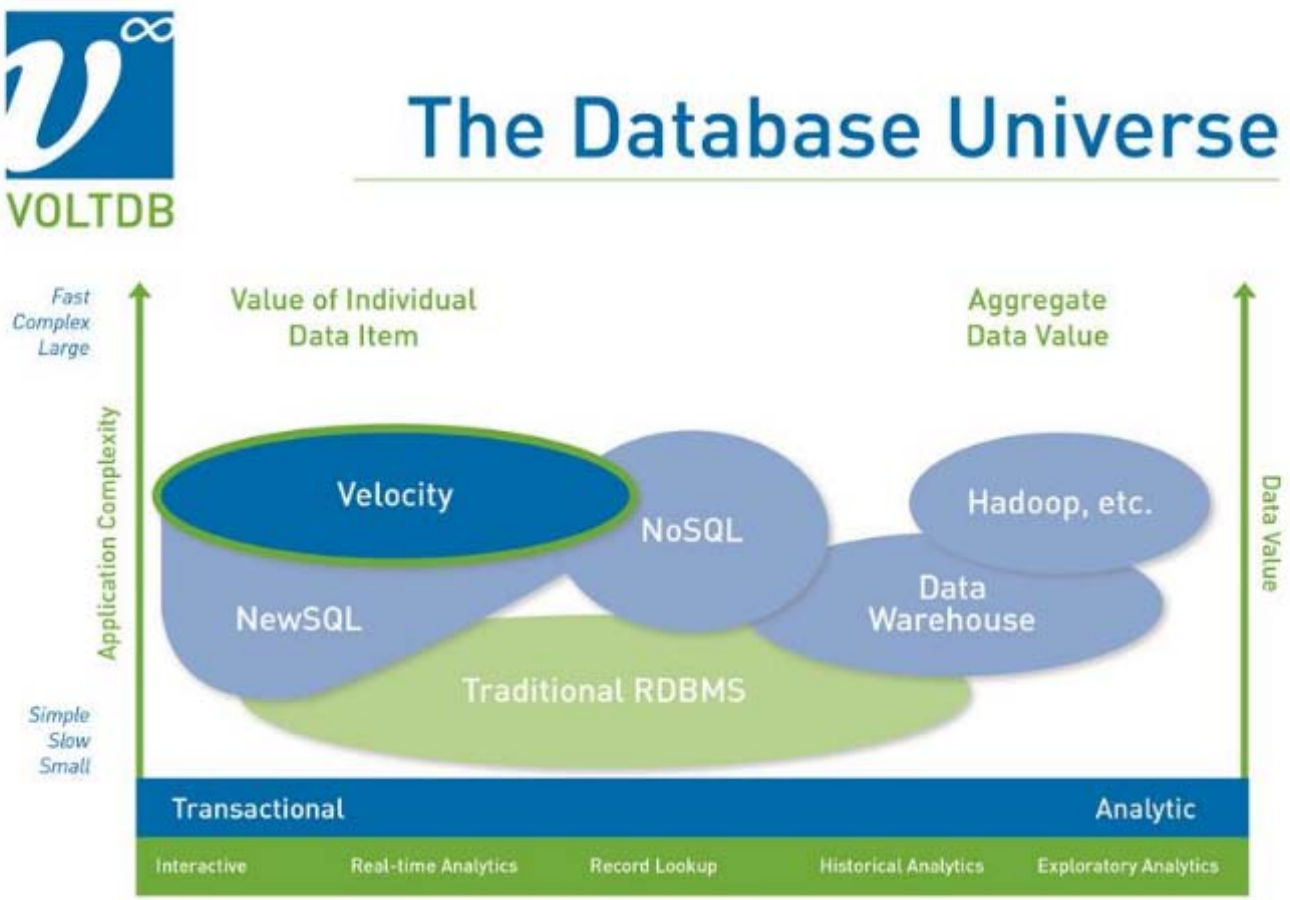
UK Data Service

A working definition of Big Data

Data sets that exceed the boundaries and sizes of normal processing capabilities, forcing you to take a non-traditional approach



Data and Information Products



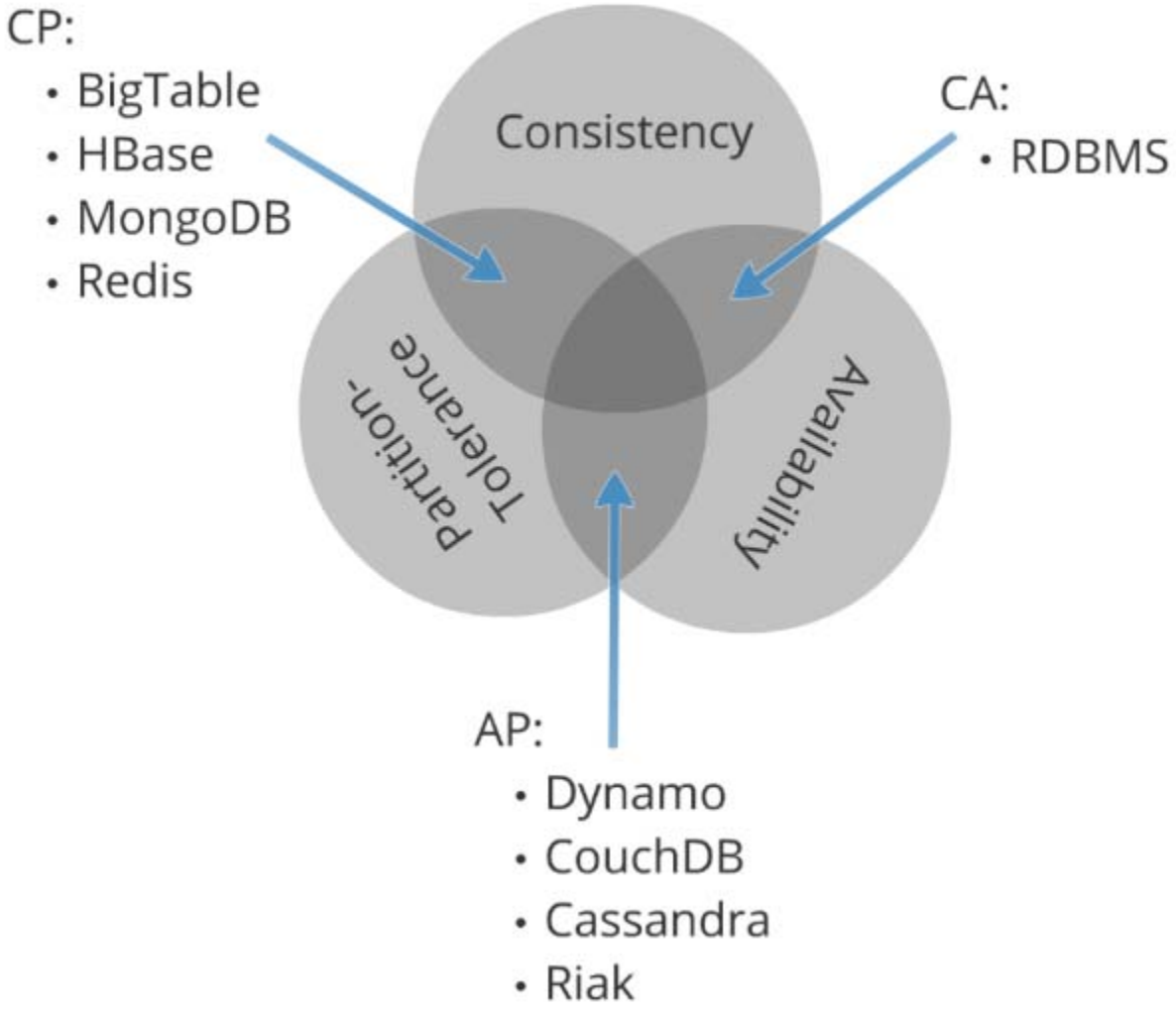
RDBMS/NoSQL/NewSQL

- RDBMS have always been distinguished by the ACID principle set (atomicity, consistency, integrity, and durability), which ensures that data integrity is preserved at all costs.
- Most NoSQL products jettison ACID performance to achieve data storage flexibility. They remove hard constraints, such as tabular row-store and strict data definitions, and they provision for scale with distributed architectures supporting high-performance throughput.
- NewSQL, retain both SQL and ACID, but they overcome the performance overhead of RDBMS caused by features such as latching shared data structures, buffer pooling, record level locking, and write-ahead logging, primarily by embracing distributed computing architectures.

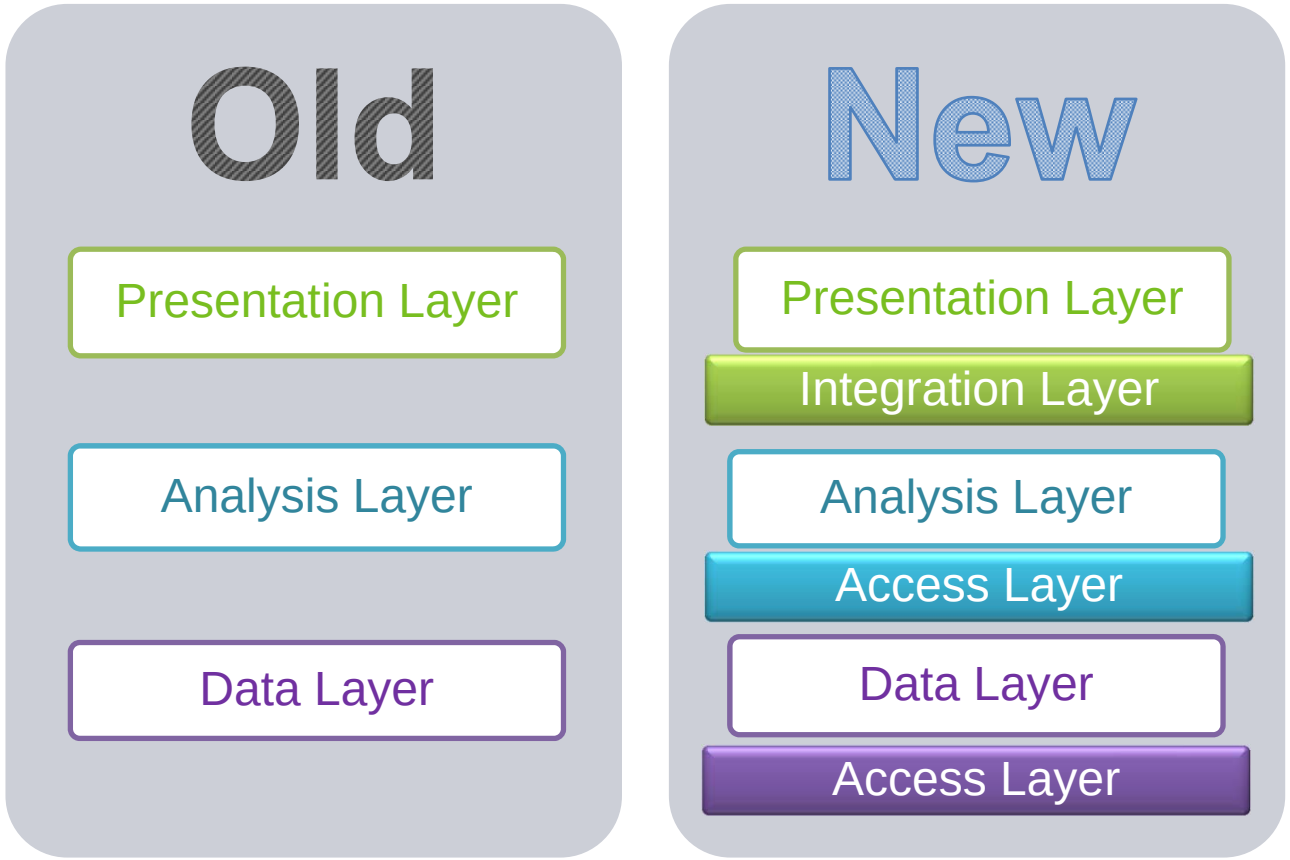
Trade-off

Characteristic	RDBMS	NoSQL	NewSQL
ACID compliance (Data, Transaction integrity)	Yes	No	Yes
OLAP/OLTP	Yes	No	Yes
Data analysis (aggregate, transform, etc.)	Yes	No	Yes
Schema rigidity (Strict mapping of model)	Yes	No	Maybe
Data format flexibility	No	Yes	Maybe
Distributed computing	Yes	Yes	Yes
Scale up (vertical)/Scale out (horizontal)	Yes	Yes	Yes
Performance with growing data	Fast	Fast	Very Fast
Performance overhead	Huge	Moderate	Minimal
Popularity/community Support	Huge	Growing	Slowly growing

Service Providers



Data as a service – unified architecture

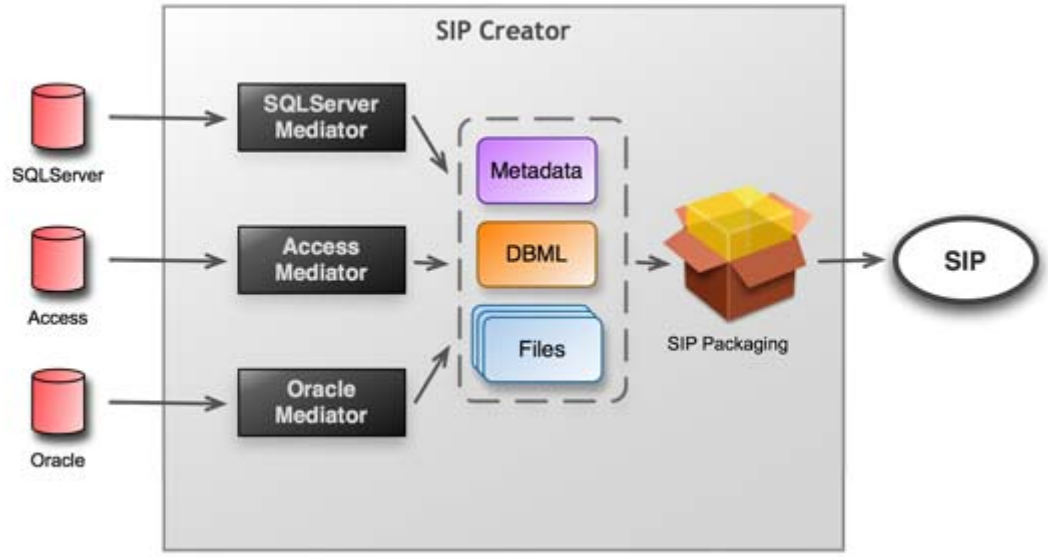


Building the new layers with APIs/web services to integrate with existing archive services

UK Data Service

Key challenge:
Preservation of Executable Content

- We need to establish a robust ecosystem for long-term preservation of software, RDMS, and other executable content.



- Olive Archive https://olivearchive.org/about/ Carnegie Mellon University / IBM

UK Data Service

Making new solutions – the Web of Data

The Evolving Web Paradigm

by jeffsayre.com

The Data Space

	Web 1.0	Web 2.0	Web 3.0
Data State	Primarily read only; data static; indirect sharing through hyperlinks	Read and write; data shareable; content interactive; data mashable	Read, write, execute; dynamic Web services; data linked and structured; data meshable
Data Storage	Data heavily locked into operating environment; mainly flat file with some relational; most data centralized	Mainly RDBMS with some OODBMS (even ORDBMS); some decentralization; most data still not abstracted, instead locked to operating environment	RDBMSs used more strategically with Document and Graph databases coming into their own; NoSQL DBs become popular; emergence of a Global Meta-DBMS
Data Discovery	Via rudimentary text search	Via advanced text search but most data closed to global querying	Via semantic search and query; data fully abstracted with heavy reliance on dereferenceable URIs; data stored in disparate locations are globally queryable, can be integrated into a federated DB
Data Relevance	Low	Medium	High, very targeted

Questions

Nathan Cunningham

Functional Director

Big Data Network Support

njcunna@essex.ac.uk

http://ukdataservice.ac.uk