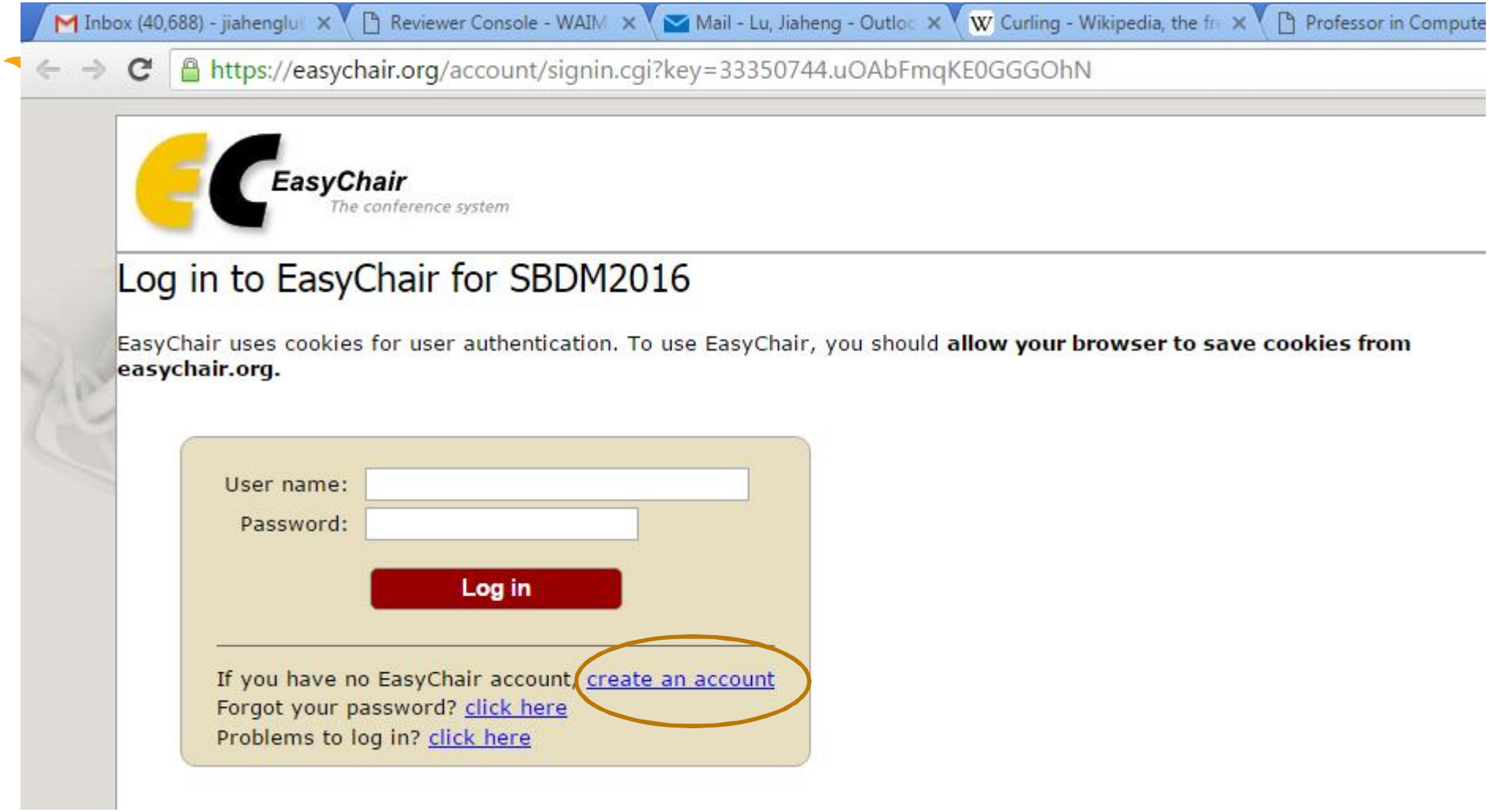
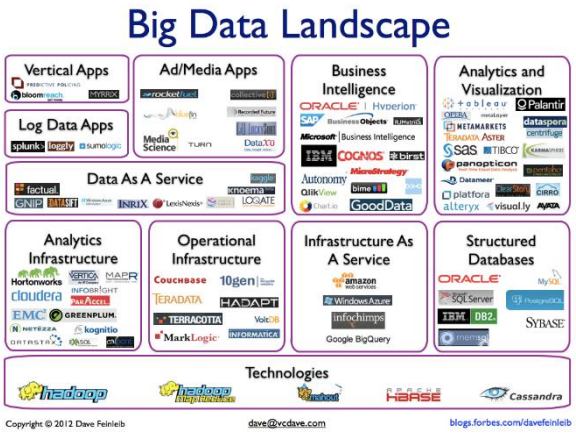
This seminar will use the EasyChair system to manage the reports and reviews.

The submission website is https://easychair.org/conferences/?conf=sbdm2016

The deadline of the first version of the report is 7 Mar, 2016.


Inbox (40,688) - jiahenglu
Reviewer Console - WAIM
Mail - Lu, Jiaheng - Outlo
Curling - Wikipedia, the fr
Professor in Compute
https://easychair.org/account/signin.cgi?key=33350744.uOAbFmqKE0GGGOhN
EasyChair
The conference system
Log in to EasyChair for SBDM2016
EasyChair uses cookies for user authentication. To use EasyChair, you should allow your browser to save cookies from easychair.org.
User name:
Password:
Log in
If you have no EasyChair account, create an account
Forgot your password? click here
Problems to log in? click here



Information when you prepare your presentation

Seminar on big data management

Lecturer: Jiaheng Lu

Spring 2016

Your attitudes are important

- Are you INTERESTED in your topic?
  - If no, get a different one!
  - If yes, ACT LIKE IT
- If YOU are not excited …
  - Can not expect OTHER people to be!

Avoid dead man talking

- Are you hiding behind the podium?
- Are your hands/face motionless?
- Are you staring …
  - at your laptop
  - At the screen
  - At the ceiling
- IF SO … you are probably BORING

Is your back to the audience?


9. Do not read directly from your notes.
Keep eye contact with your audience.

Right here. See ?

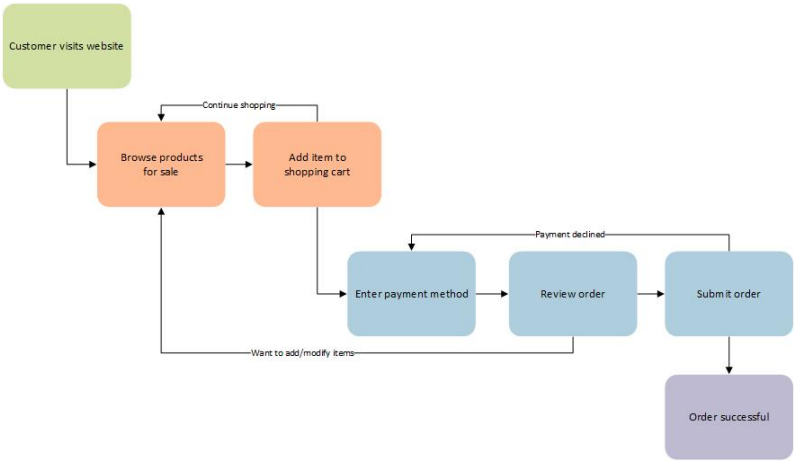
- Do not point at your laptop screen
- They cannot see it

Practice is important

- Practice makes perfect (at least three times before your talk)
- Do not read your slides like a script
- Most people lose 20 IQ points in front of an audience

Picture the idea

- In general, do not have only text on most of your slides
- Try to draw diagrams wherever applicable
- (Well-drawn) pictures easier to understand

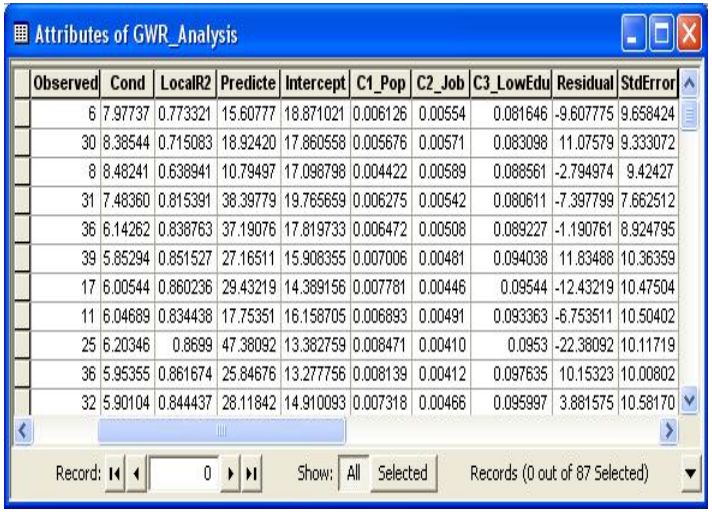


Anticipatory lecturing

- Do not be a tease
- Let the audience think at their own pace
- It only provides benefit if there is a ”surprise” result.

Results explanation

- You have lots of cool results
  - No one can read this
  - No one can undstand this
- Graphs are your friend



Attributes of GWR_Analysis

Observed	Cond	LocalR2	Predicte	Intercept	C1_Pop	C2_Job	C3_LowEdu	Residual	StdError
6	7.97737	0.773321	15.60777	18.871021	0.006126	0.00554	0.081646	-9.607775	9.658424
30	8.38544	0.715083	18.92420	17.860558	0.005676	0.00571	0.083098	11.07579	9.333072
8	8.48241	0.638941	10.79497	17.098798	0.004422	0.00589	0.088561	-2.794974	9.42427
31	7.48360	0.815391	38.39779	19.765659	0.006275	0.00542	0.080611	-7.397799	7.662512
36	6.14262	0.838763	37.19076	17.819733	0.006472	0.00508	0.089227	-1.190761	8.924795
39	5.85294	0.851527	27.16511	15.908355	0.007006	0.00481	0.094038	11.83488	10.36359
17	6.00544	0.860236	29.43219	14.389156	0.007781	0.00446	0.09544	-12.43219	10.47504
11	6.04689	0.834438	17.75351	16.158705	0.006893	0.00491	0.093363	-6.753511	10.50402
25	6.20346	0.8699	47.38092	13.382759	0.008471	0.00410	0.0953	-22.38092	10.11719
36	5.95355	0.861674	25.84676	13.277756	0.008139	0.00412	0.097635	10.15323	10.00802
32	5.90104	0.844437	28.11842	14.910093	0.007318	0.00466	0.095997	3.881575	10.58170

Record: 0 Show: All Selected Records (0 out of 87 Selected)

Keep it simple

- Do you really need all those equations?
  - This is very instance-dependent
  - Depends on what you are discussing
  - Depends on your audience
- Sometimes you may need them
  - Explain the variables

    and what they mean
  - Given a ”plain-text” description of it

$$\nabla \cdot \mathbf{A} = \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z}$$

$$= (\text{rate of change of } \mathbf{A} \text{ in x-direction}) + (\text{rate of change of } \mathbf{A} \text{ in y-direction}) + (\text{rate of change of } \mathbf{A} \text{ in z-direction})$$

Summay and conclusion

- Remember to summarize work and results
- Giving ”selling” points here
  - 30X performance increase with only 10% area penalty

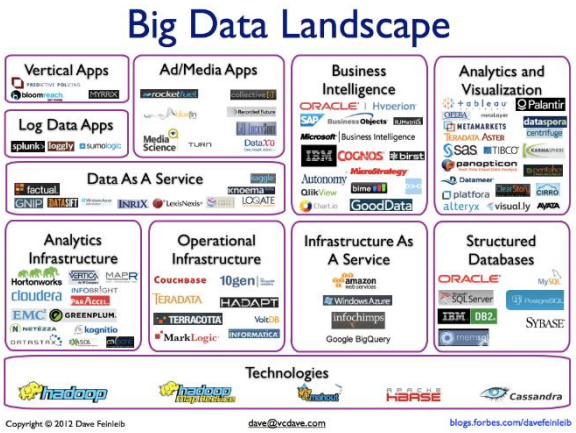
Before your presentation

- Please discuss with me about your presentation paper at least one week in advance of your presentation

- Please send the slides to the opponent and me before your presentation

Bad presentations

Wish your presentation to be a good presentation



New trends of big data management in 2016

Seminar on big data management

Lecturer: Jiaheng Lu

Spring 2016

Trend 1: Spark grows fast

Spark “on the radar”

- 2008 - Yahoo! Hadoop team collaboration with Berkeley Amp/Rad lab begins
- 2009 - Spark example built
- 2011 - “Spark is 2 years ahead of anything at Google”
  - Conviva seeing good results w Spark
- 2012 - Yahoo! working with Spark / Shark
- Today - Many success stories
  - Early commercial support

Spark updates Hadoop

- Hardware had advanced since Hadoop started:
  - Very large RAMs, Faster networks (10Gb+)
  - Bandwidth to disk not keeping up
- MapReduce awkward for key big data workloads:
  - Low latency dispatch (E.G. quick queries)
  - Iterative algorithms (E.G. ML, Graph…)
  - Streaming data ingest

Spark, “lingua franca?”

- Support for many development techniques
  - SQL, Streaming, Graph & in memory, MapReduce
  - Write “UDFs” once and use in all contexts
- Small, simple & elegant API
  - Easy to learn and use; expressive & extensible
  - Retains advantages of MapReduce (fault tolerance…)

Spark often better

- Today you will hear many success stories from teams who have converted Hadoop based workloads to Spark and seen:
  - Huge speedups and Big cost savings
- But there do exist cases were Hadoop is superior…
  - Proven to work at the largest scales
  - Mature & widely commercially supported
  - Much larger ecosystem of solutions and tools

Spark complements Hadoop

- Spark leverages Hadoop ecosystem
  - HDFS, HCatalog, Data Input/OutputFormats
  - Huge investment in data collection & tooling

Spark the “lingua franca”

- Data scientists & Developers need an open standard for sharing their Algorithms & functions, an “R” for big data.

- Spark best current candidate:
  - Open Source - Apache Foundation
  - Expressive (MR, iteration, Graphs, SQL, streaming)
  - Easily extended & embedded (DSLs, Java, Python…)

Trend 2: most operational DBMSs will offer multiple data models, relational and NoSQL, in a single DBMS platform.

Operational databases

- Operational database management systems are used to manage dynamic data in real-time.

- Operational databases use NoSQL DBMS engines

and distributed database architecture that provides high availability and fault tolerance

Gartner Magic quadrant for operational database management systems

- By 2017, all leading operational DBMSs will offer multiple data models, relational and NoSQL, in a single DBMS platform.

Using SQL to query NoSQL and relational databases

- SQL: SELECT * FROM NoSQL WHERE category='NoSQL‘ (Support by CouchBase)

- JSON results:


```
{
  "name": "Couchbase Server",
  "version": "4.0",
  "category": "NoSQL",
  "features": [ "name": "N1QL", "capabilities": ["JOIN", "NEST", "UNNEST“ ] ]
}
```

Using SQL to query NoSQL and relational databases

- SQL: SELECT * FROM RDB WHERE category='RDB'

- Relational results:

Name	Version	category	Features_name	Features_capabilities
MySQL	4.0	RDB	SQL	JOIN, NEST, GROUP-BY

Using SQL to join NoSQL and relational databases

- SQL: SELECT * FROM RDB, NoSQL WHERE RDB.name= NoSQL.name

- Relational results:

Name	Version	category	name	capabilities
ORACLE	1.0	NoSQL	SQL	JOIN

JSON results:


```
{
  "name": “ORACLE",
  "version": “12c",
  "category": “RDB",
  "features": [ "name": “SQL",
"capabilities": ["JOIN" ] ]
}
```

Trend 3: Self-service data preparation tools are exploding

Self-service data preparation

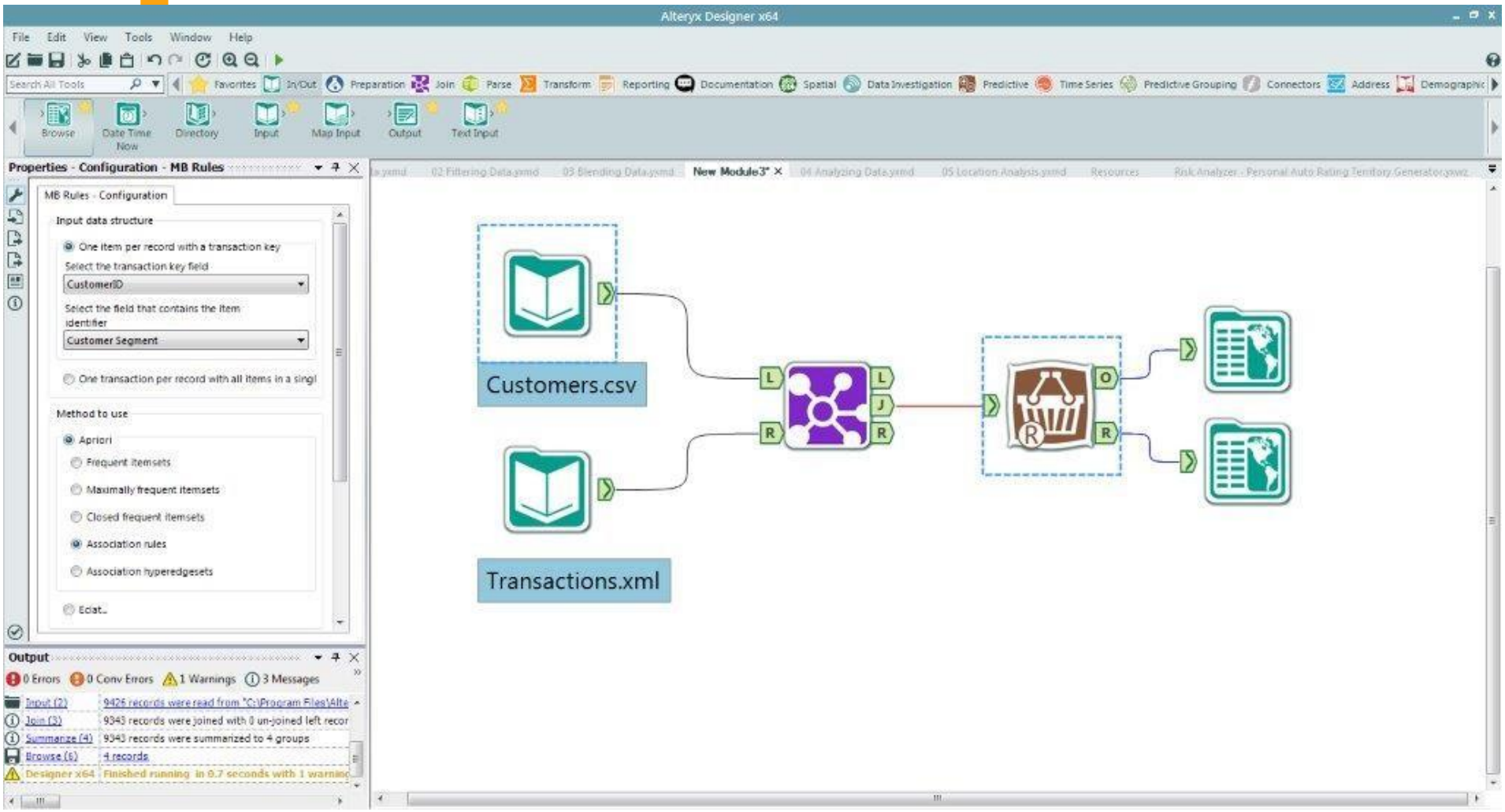
- Business users want to reduce the time and complexity of preparing for analysis big data

- Products: Alteryx, Trifacta, Paxata

Alteryx

- Alteryx is an American computer software company based out of Irvine, California.
- The company's products are used for data blending and advanced data analytics.
- Alteryx has a stated goal of enabling advanced analytics to be performed by non-specialists

Alteryx workflow



Trifacta

- Trifacta is a data transformation platform provider that enables business analysts, data scientists and IT programmers to transform data into a usable form for analysis.

TRIFACTA
PEOPLE DATA COMPUTATION

Paxata

- Paxata develops self-service data preparation software
- Paxata's software is used to combine data from different sources, then check it for data quality issues, such as duplicates and outliers.
- Algorithms and machine learning automate certain aspects of data

Summary

- Three trends for big data management in 2016:

1. Spark grows fast and more popular
2. One DBMS will host NoSQL and SQL
3. Popularity of self-service data preparation tools will explode.