Architecting for Data Science

O'Reilly Software Architecture Conference Boston, March 19, 2015

> Johann Schleier-Smith CTO, if(we)

johann@ifwe.co github.com/ifwe

@jssmith



ef9053e4c5a713d0814a14947f216b4a1e7333bc,1214981,abcat0515039,"Logitech speakers","2011-08-11 09:03:58.306","2011-08-11 09:02:17.05 fa4bbed0c6a651c74db4eda0f523da8ff869e569,2573486,pcmcat218000050001,Asus,"2011-08-11 09:04:11.288","2011-08-11 09:00:34.853" e831860426098c8f9191318f71c56d05b7039d21,2496342,pcmcat209400050001,Xperia,"2011-08-11 09:04:11.796","2011-08-11 09:03:06.493" 29a921902ab04b3dc339f4b423cf824ea2869ec3,9097764,abcat0410018,10-20,"2011-08-11 09:04:14.571","2011-08-11 09:03:19.574" e6e83ccc269dc26b049f231139f52a1228b0b76c,1686194,abcat0703002,"Mass effect 2","2011-08-11 09:04:46.452","2011-08-11 09:04:19.712" 1bccd5893543a9324589a13b85286cd9af150121,2627475,cat02716,"Maybach music","2011-08-11 09:04:56.875","2011-08-11 09:04:41.844" d363ed695b34ef4e05695f9b81431b5e24757426,1232769,abcat0208011,"ipod speaker dock","2011-08-11 09:05:11.381","2011-08-11 09:02:01.21 f1c0aa8e2ecb1bb3e7cd87a8c0405557eb1fae10,9824298,pcmcat230800050019,"Turtle beach headset","2011-08-11 09:05:16.785","2011-08-11 09 c57fec4bb43bcfd2effbabe5a61c1f13cb52a06c,9835567,abcat0208031,Isimple,"2011-08-11 09:05:21.818","2011-08-11 09:05:02.714" a0ebda10ad523306a6830af4eef06482acf48cc2,9420361,pcmcat183800050007,"computer charger","2011-08-11 09:05:38.067","2011-08-11 09:04: 022969be550859b88fd48b4d3ce22ce443b4c434,1816383,pcmcat170900050018,Webcam,"2011-08-11 09:05:40.698","2011-08-11 09:02:39.729" c57fec4bb43bcfd2effbabe5a61c1f13cb52a06c,8760931,abcat0307018,Isimple,"2011-08-11 09:06:35.152","2011-08-11 09:05:02.714" 9a113240869815c067a369ecf097c9ed5549d5ba,3168049,cat02719,jay-z,"2011-08-11 09:06:39.825","2011-08-11 09:06:23.153" 1bccd5893543a9324589a13b85286cd9af150121,2965038,cat02716,"Gucci mane","2011-08-11 09:06:46.815","2011-08-11 09:05:50.867" 46257e80a10205186c0ff532aad27b1ea079401d,2817743,pcmcat209000050008,Galaxy,"2011-08-11 09:06:48.036","2011-08-11 09:05:56.298" c5658ce9de1bae04b6472ab77c9f86a1cb00524a,1323166,abcat0508026,"Video editing software","2011-08-11 09:06:49.754","2011-08-11 09:05: 1e807283c606049bf18bef5f25 bc2fc4 9ee 5 30 - 49 - 5 cat2120 0050008,"' o dual corre "2011 58-11 5007' 6067","2011-08-11 09:04:34.5 90bfc8 2b5f41 240 105,ab) t0101005 - 5 na (nic gi 01 "201- 5-11 01 07:53 87" '2011-08-11 09:06:53.312" 5de607 80 dez 110 09 pc1 at1435100500 1 'Skullo no tow', 2011 0 11 09 7.22.77","2011-08-11 09:06:32. 907 dat 25 110 09 pc1 at1435100500 1 'Skullo no tow', 2011 0 11 09 7.22.77","2011-08-11 09:06:32. d8c14c4a46793addd1d0f11e32 4cde5c0acecab3dfe51bb260e3 953dc1bb1c9d5145a50ffae5a6 4113ac371d0c1964156065da7d 1e807283c606049bf18bef5f2572a8c15505d1db,3168049,cat02719,"watch the throne","2011-08-11 09:07:26.639","2011-08-11 09:06:49.087" d311a3524a418b3349d418acc1332a250abcca8a,8229044,abcat0503013,cables,"2011-08-11 09:07:26.871","2011-08-11 09:05:58.304" 90b1b7d9a5c27a98428901735f255a12ec4afb71,2140037,abcat0401004,cameras,"2011-08-11 09:07:28.346","2011-08-11 09:04:46.199" f8cbee1885e089c869e78d6d0e6c105aafc42bac,2847643,abcat0503013,"Wireless n","2011-08-11 09:07:44.65","2011-08-11 09:07:37.624" b9df0b9099c87224c3ea60eed48d47aeb53d85e4,1051329,abcat0715007,"Ps3 move","2011-08-11 09:07:51.897","2011-08-11 09:06:37.895" 723bea69e272a99ec79b818b6f983b138d1a8ab5,9225377,abcat0201011,"iPod nano","2011-08-11 09:07:57.449","2011-08-11 09:06:58.583" 9f7fbe852db0d845d2bf43472c4f4b54490c861b,8280834,abcat0107004,"hdtv antenna","2011-08-11 09:08:11.619","2011-08-11 09:06:49.658" 8e7d1caf3c3fd56bee3697f63188be61c630048a,3048064,pcmcat247400050000,"Lap tops","2011-08-11 09:08:26.366","2011-08-11 09:07:51.508" 3920405da68ea703996bc425893e3f9d2aa41648,2128267,abcat0801003,"virgin mobile","2011-08-11 09:08:36.533","2011-08-11 09:07:11.16" b3d51f75dc3e595cbf5871bae9c8d5b0656aa51c,2408109,abcat0101001,"65 inch tv","2011-08-11 09:08:48.206","2011-08-11 09:07:24.274" 3658871850982b856204b9a338127ef8a110b0dc,9952217,pcmcat151600050006,"Fender starcaster","2011-08-11 09:09:00.872","2011-08-11 09:08 c9544776252f9334afea7416a0df54df4fe08eed,9492426,pcmcat143000050007,"beats by dre","2011-08-11 09:09:01.29","2011-08-11 09:08:12.56 90b1b7d9a5c27a98428901735f255a12ec4afb71,2642464,abcat0401004,cameras,"2011-08-11 09:09:03.367","2011-08-11 09:04:46.199" f8cbee1885e089c869e78d6d0e6c105aafc42bac,7898082,abcat0503003,"Wireless n","2011-08-11 09:09:14.413","2011-08-11 09:07:37.624" bd0add05492ec9301c902345c57bed20d8c303bc,2715258,pcmcat247400050000,intel,"2011-08-11 09:09:16.892","2011-08-11 09:07:08.29"



c5c73ded3bbe43f9baddd220d0461e89a042fa6e,2822239,abcat0307015,auxiliary,"2011-08-11 09:09:18.47","2011-08-11 09:06:38.533"



Harvard Business Review



DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

value from data

Alternative Definitions

making discoveries in the world of big data

statistics + machine learning + scalable computation + visualization + computer science + business acumen + skilled communication

extraction of knowledge from data



Related and Alternative Language data mining

business intelligence

predictive modeling

statistics

knowledge extraction

forecasting

business reporting

analytics

Value





understanding

predictions

revenue

customer satisfaction

Types of Value product improvements

projections

new inspirations



Today's Examples

Recommendation engine for dating product

- >10 million candidates to draw from >1000 updates/sec
- Must be responsive to current activity
- Users expect instant query results



- Real-time is challenging
- Human behavior is complicated, especially in social context
- Previous interactions are perhaps our best hope for predicting future interactions



Value

- Human connections
- User engagement ecosystem
- Subscription and other revenues





Kaggle competition with Best Buy data

https://www.kaggle.com/c/acm-sf-chapter-hackathon-small





Kaggle competition with Best Buy data













Sah J LT







"outgoing and social (heavy messaging --especially distant recipients and opposite gender, many outgoing comments, many friend requests to distant people), doesn't play Pets much"

"heavy user overall, (pets, meet me, messaging)"

"receives many messages, active user, views many profiles, doesn't use meet me, sends many messages to distant people"

Value





计算用 医沙根 劉氏 医骨骨骨骨炎 法法律法院 网络拉拉科教教学科科教教教教教教教教教教教教 <1d>cat02001<//d> 应该目下后该多的总领主你们没不少资源目前的信息从平平111月1日也会从时间下的C自由出口性生活; <name>Music</name> 王又臣商务团任商主任关于王又臣商务 《日臣帝国务局党财务局主主政务关プ日臣官务 </category> 金玉翁 医毛肉肉毛肉 医白霉素 网络含金玉 的复数形式 使自主自己的复数形自主的 医 <category>

 Inewsfalses/news
 calegory>
 <id>calegory>
 <id>calegory></ <id>cat02506</id> <name>Easy Listening </categoryPath> <lists/> (6月5月163年有16亩水76月均有1亩3 314131631631660000047163 <customerReviewCount/> name:Movies & MusikeusterReviewAveragez> BETERMITERETERIZE STORESTERIES STORE <format>CD</format> anvertisedErteekestrictentekeekekerigtsetiedvertisedErteeshipphightruetzfreeShippingx minimumBigplayPrice/> <id>cat02001x/1dx <inStoreAvailability>false</inStoreAvailability> 王帝原来民主任王帝和教师主命民任王帝的法法的意义的意思。因此是他的意义的问题,我们还是在这些法 <inStoreAvailabilityText>Store Pickup: Not Availabl 10F1V目前23 <inStoreAvailabilityTextHtml>Store Pickup; lerstartbate/+ <category> <inStoreAvailabilityUpdateDate>2012-07+02T23:06:52<</pre> 10F目前目前102506371133 <itemUpdateDate>2012-07-28T10:28:36</itemUpdateDate 41日1日1日前13下前1665/01日日1日13/CaleBory》 <onlineAvailabilityText>Shipping: Usually leaves ou 的目标的同时的自然目的主要自然分白目前的所有自我在自己的工具自我的发 <onlineAvailabilityTextHtml>Shipping;</ OULIGENEER/> Stists/> <onlineAvailabilityUpdateDate>2012-07-02T23:06:52 <releaseDate>1994-05-24</releaseDate> TFEEDUEDELVFUFEDBSEDWFEBStomerRev1ewAverBEe/> <shippingCost>0.00</shippingCost>

 xcrusssell/s
 <format>CASSETTE</format> <shipping>

 xsalesRankShortTerm/s<freeShipping>false</freeShipping>false<//treeShipping>dimd>0.00</pround>

 xsalesRankMediumTerm/sinStoreAvailability>false<//d>

 xsalesRankLongTerm/s
 <inStoreAvailabilityText>StoreActionaly>

 xsalesRankLongTerm/s
 <inStoreAvailabilityText>StoreActionaly>

 xsalesRankLongTerm/s
 <inStoreAvailabilityText>StoreActionaly>

 xsalesRankLongTerm/s
 <inStoreAvailabilityText>StoreActionaly>

 xsalesRankLongTerm/s
 <inStoreAvailabilityText>StoreActionaly>

 xsalesRankLongTerm/s
 <inStoreAvailabilityText>StoreActionaly>

 xsalesRankLongTerm/s
 <inStoreAvailabilityTextHtml>

 xsalesRankLongRank/s
 <inStoreAvailabilityTextHtml>



Dating product data

- Vote history
- Social interaction history
- Profile information





"timestamp": "2011-10-31 09:48:46", 'query' : 'Assassin's Creed', "skuSelected" : "2670133"



product views





Formats for Data

log files

xml files

web services

relational databases

unstructured documents

spreadsheets

technical data

reference data

government data



Ivpes of Data

usage records

sensor data

academic data

yet uncollected data



<u>Vasant Dhar. 2013. Data science and prediction. Commun. ACM 56, 12 (December 2013), 64-73.</u> And International Telecommunication Union (ITU) and United Nations Population Division via <u>www.internetlivestats.com/internet-users/</u>

data quantity

data variety

data velocity

Trends

machine learning maturing



Machine Learning supervised methods

classification

what matters most is mapping of data

clustering unsupervised methods

decision trees



Machine Learning Techniques

- Classification (Logistic Regression, Decision) Trees, Random Forests)
- Prediction (Generalized Linear Models, Support Vector Regression, ...)
- Clustering (K-Means, Hierarchical, Latent Dirichlet Allocation, ...)
- Collaborative filtering, ...

Features often matter more than choice of algorithm

计算用 医沙根 劉氏 医骨骨骨骨炎 法法律法院 网络拉拉科教教学科科教教教教教教教教教教教教 <1d>cat02001<//d> 应该目下后该多的总领主你们没不少资源目前的信息从平平111月1日也会从时间下的C自由出口性生活; <name>Music</name> 王又臣商务团任商主任关于王又臣商务 《日臣帝国务局党财务局主主政务关プ日臣官务 </category> 金玉翁 医毛肉肉毛肉 医白霉素 网络含金玉 的复数形式 使自主自己的复数形自主的 医 <category>

 Inewsfalses/news
 calegory>
 <id>calegory>
 <id>calegory></ <id>cat02506</id> <name>Easy Listening </categoryPath> <lists/> (6月5月163年有16亩水76月均有1亩3 314131631631660000047163 <customerReviewCount/> name:Movies & MusikeusterReviewAveragez> BETERMITERETERIZE STORESTERIES STORE <format>CD</format> anvertisedErteekestrictentekeekekerigtsetiedvertisedErteeshipphightruetzfreeShippingx minimumBigplayPrice/> <id>cat02001x/1dx <inStoreAvailability>false</inStoreAvailability> 王帝原来民主任王帝和教师主命民任王帝的法法的意义的意思。因此是他的意义的问题,我们还是在这些法 <inStoreAvailabilityText>Store Pickup: Not Availabl 10F1V目前23 <inStoreAvailabilityTextHtml>Store Pickup; lerstartbate/+ <category> <inStoreAvailabilityUpdateDate>2012-07+02T23:06:52<</pre> 10F目前目前102506371133 <itemUpdateDate>2012-07-28T10:28:36</itemUpdateDate 41日1日1日前13下前1665/01日日1日13/CaleBory》 <onlineAvailabilityText>Shipping: Usually leaves ou 的目标的同时的自然目的主要自然分白目前的所有自我在自己的工具自我的发 <onlineAvailabilityTextHtml>Shipping;</ OULIGENEER/> Stists/> <onlineAvailabilityUpdateDate>2012-07-02T23:06:52 <releaseDate>1994-05-24</releaseDate> TFEEDUEDELVFUFEDBSEDWFEBStomerRev1ewAverBEe/> <shippingCost>0.00</shippingCost>

 xcrusssell/s
 <format>CASSETTE</format> <shipping>

 xsalesRankShortTerm/s<freeShipping>false</freeShipping>false<//treeShipping>dimd>0.00</pround>

 xsalesRankMediumTerm/sinStoreAvailability>false<//d>

 xsalesRankLongTerm/s
 <inStoreAvailabilityText>StoreActionaly>

 xsalesRankLongTerm/s
 <inStoreAvailabilityText>StoreActionaly>

 xsalesRankLongTerm/s
 <inStoreAvailabilityText>StoreActionaly>

 xsalesRankLongTerm/s
 <inStoreAvailabilityText>StoreActionaly>

 xsalesRankLongTerm/s
 <inStoreAvailabilityText>StoreActionaly>

 xsalesRankLongTerm/s
 <inStoreAvailabilityText>StoreActionaly>

 xsalesRankLongTerm/s
 <inStoreAvailabilityTextHtml>

 xsalesRankLongRank/s
 <inStoreAvailabilityTextHtml>



tools of

trace





- Created in 1993
- Implementation of S language but also inherits from Scheme
- Object oriented code is possible but not encouraged
- Vast high-quality package ecosystem
- Data is vectors and *data frames*





Demo

R Studio

```
💽 - 🔂 - 🔒 🔒 🖾 Go to file/function
  Untitled1* x mydata x
         📙 🗌 Source on Save 🛛 🔍 🖉 🖌 📃
                                                 Run 时 Source 🗸
  0
      mydata <- read.csv("~/demo/data-1.csv")</pre>
   1
      plot(mydata)
   2
      attach(mydata)
   3
      m <- lm(Y \sim X)
      summary(m)
   5
       abline(m)
   6
         🗾 (Top Level) 🗘
   2:1
  Console ~/ 📣
 > plot(myData)
 > par(mfrow(c(1,1)))
 Error in par(mfrow(c(1, 1))) : could not find function "mfrow"
 > par(mfrow=(c(1,1)))
 > plot(myData)
 > attach(myData)
 The following objects are masked from myData (pos = 3):
     Χ, Υ
 > plot(X~Y)
 > m <= lm(X~Y)
 Error in m \ll lm(X \sim Y) : comparison of these types is not implemented
 > m <- lm(X~Y)
 > summary(m)
 Call:
 lm(formula = X \sim Y)
 Residuals
```





- Statistics
- Visualization
- Machine learning
- REPL, scripts, interactive IDE
- In-memory data sets

puthon





IP[y]: IPython Interactive Computing








http://scikit-learn.org/stable/auto_examples/linear_model/plot_iris_logistic.html







m

- Sophisticated machine learning libraries

- More of a general purpose language than R Arrays and matrices as basic data structures Supports data frames through Pandas
- Generally limited to in-memory data sets



- Leverages commodity hardware to store large data sets at low cost
- Vibrant and diverse ecosystem
- Popular but not always best solution
- Probably best viewed as marketing terminology, as opposed to technology





130 freely licensed open source projects listed in the Hadoop Ecosystem Table

https://hadoopecosystemtable.github.io/

Category	Number of projects
Distributed Filesystem	7
Distributed Programming	18
NoSQL Database	
Document Data Model	3
Stream Data Model	
Key-Value Data Model	Z
Graph Data Model	G
NewSQL	Ç
SQL-On-Hadoop	
Data Ingestion	11
Service Programming	7
Scheduling	C
Machine Learning	6
Benchmarking	5
Security	3
System Deployment	12
Applications	5
Development Frameworks	2
Categorize Pending	16







Hadoop for Data Scientists

Pulling data from repository (SQL, Hive)

 MapReduce programming (Java, Scala, Pig, Python)

 Spark in-memory framework is gaining adoption rapidly

tools rarely used in data science

version control

shared code

agile methodology

automated testing

automated deployment

code review

software architecture



the cycle of data science







ef9053e4c5a713d0814a14947f216b4a1e7333bc,1214981,abcat0515039,"Logitech speakers","2011-08-11 09:03:58.306","2011-08-11 09:02:17.05 011-08-11 09:02:39.729" 8-11 09:05:02.714" "2011-08-11 09:05:50.867" 1e807283c606049bf18bef5f2572a8c15505d1db,3168067,cat02719,"watch the throne","2011-08-11 09:07:06.472","2011-08-11 09:06:49.087" 4cde5c0acecab3dfe51bb260e3490bfc882b6f41,2408109,abcat0101001 👝 as ac 🗖 gt30","2011-08-11 09:07:13.387","2011-08-11 09:06:53.312" kullcandy low","2011-08-11 09:07:22.77","2011-08-11 09:06:32. -08-11 09:07:25.379","2011-08-11 09:05:48.273"

> 09:07:57.449"."2011-08-11 09:06:58.583" 09:08:11.619"."2011-08-11 09:06:49.658" 09:08:26.366","2011-08-11 09:07:51.508" -11 09:08:36.533","2011-08-11 09:07:11.16" inch tv","2011-08/11 09:08:48.206","2011-08-11 09:07:24.274"

er P



if(we)

- Profitable startup actively pursuing big opportunities in social apps
- Millions of users on existing products
- Thousands of social contacts per second



what it should look like

- 1. Gain understanding of the product usage
- 2. See opportunity to make the product better
- 3. Create training data
- 4. Train predictive models
- 5. Put models in production
- 6. See improvements

what it often looks like

- 1. Gain understanding of the product usage
- 2. See opportunity to make the product better
- 3. Pull records from relational database to create interesting features (usually aggregates)
- 4. Train predictive models
- 5. Go implement models for production
- 6. See improvements

- 1. Gain understanding of the product usage
- 2. See opportunity to make the product better
- 3. Pull records from relational database to create interesting features (usually aggregates)
- 4. Train predictive models
- 5. Go implement models for production
- 6. See improvements



- 1. Gain understanding of the product usage
- 2. See opportunity to make the product better
- 3. Pull records from relational database to create interesting features (usually aggregates)
- 4. Train predictive models
- 5. Go implement models for production
- 6. See improvements

Cool! Was it worth it?



implementation pain points

- Data scientist hands model description to software engineer
- May need to translate features from SQL to Java Aggregate features require batch processing May need to adjust features and model to achieve
- real-time updates
- Fast scoring requires high-performance inmemory data structures





one right way to data

one right way to data

event history

everything is an event

- Bob registers
- Alice registers
- Alice updates profile
 - Bob opens app
- Bob sees Alice in recommendations
 - Bob swipes yes on Alice
 - Alice receives push notification
 - Alice sees Bob swiped yes
 - Alice swipes yes
 - Alice sends message to Bob

architecture comparison













Production





Production





Production








Event History API

trait EventHistory def publishEvent(e: Event)

def getEvents (startTime: Date, endTime: Date, eventFilter: EventFilter, eventHandler: EventHandler

Event History API

trait EventHistory { def publishEvent(e: Event)

def getEvents(startTime: Date, endTime: Date, eventFilter: EventFilter, eventHandler: EventHandler

Event History API

trait EventHistory {
 def publishEvent(e: Event)

def getEvents(startTime: Date, real-lime endTime: Date, streaming eventFilter: EventFilter, eventHandler: EventHandler

training data comparison







Production







- 1. Gain understanding of the product usage
- 2. See opportunity to make the product better
- 3. Create training data
- 4. Train predictive models
- 5. Put models in production
- 6. See improvements

Fost cvcles!!





Live Demo

https://github.com/ifwe/antelope

https://www.kaggle.com/c/acm-sf-chapter-hackathon-small

- proprietary platform
- Not ready scale or production, but useful for demonstration purposes
- Seeking collaborators

Open source implementation derived from if(we)'s

product update events

"timestamp": "2012-05-03 6:43:15", "eventType" : "ProductUpdate", "eventProperties" : { "sku": "1032361", "regularPrice": "19.99", on it..."

```
"name" : "Need for Speed: Hot Pursuit",
"description" : "Fasten your seatbelt and
get ready to drive like your life depends
```

product view events

"timestamp": "2011-10-31 09:48:46", "eventType" : "ProductView", "eventProperties" : { "query": "Modern warfare", "skuSelected" : "2670133"

Try it yourself, code and instructions at: https://github.com/ifweco/antelope/blob/master/doc/demo.md

demo

class TermPopularity extends Feature[SC] { val ct = s.counter(queryTerms, skuViewed)

override def score(implicit ctx: SC): (Long) => Double = { val terms = ctx.query.normalize.split(" ") id => terms.map { term => ct(term, id) div ct(term) }.product

}

class TfIdf extends Feature[SC] { val termFreq = s.counter(productNameUpdatedTerms, skuUpdated) val docFreg = s.set(productNameUpdatedTerms, skuUpdated) val docs = s.set(skuUpdated)

val terms = ctx.query.normalize.split(" ") val n = docs.size() id => terms.map { term => val tf = termFreq(term, id) val df = docFreq.size(term) sqrt(tf) * sq(1D + log(n/(1D+df))) }**.**sum

```
override def score(implicit ctx: SC): (Long) => Double = {
```

```
Deviance Residuals:
            1Q Median 3Q
   Min
                                     Max
-8.4904 -0.3565 0.0000
                          0.0030
                                 2.3765
Coefficients:
           Estimate Std. Error z value Pr(>|z|)
(Intercept) 7.15236
                      0.30019 23.83 <2e-16 ***
feature 1 2.01115 0.04070 49.41 <2e-16 ***
feature 2 11.95880 0.61440 19.46 <2e-16 ***
feature 3 3.59281 0.07072 50.80 <2e-16 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
   Null deviance: 55452 on 39999 degrees of freedom
Residual deviance: 13978 on 39996 degrees of freedom
AIC: 13986
Number of Fisher Scoring iterations: 12
>
> # save output to text file
> write(as.vector(coef(mylogit)[-c(1)]) / as.vector(colsd),
   file='r_logit_coef.txt', sep=",", ncolumns=1000)
Johanns-MacBook-Pro-2:training_small johann$
```

	г				-	🔲 ant	elope — ec2-us
ec2-user@ip-172-30-0	-89:~ 775	ec2-user@i	p-17/spar	k-1.1.0/ec2		bash	8
nrograss	275	60					
progress	200	66					
progress	200	60					
progress	290	66					
progress	290	00					
progress	200	66					
progress	210	00					
progress	510 510	00					
progress	220	00					
progress	220	66					
progress	220	00					
progress	330 225	00					
progress	270	00					
progress	240	00					
progress	250	00					
progress	250	00					
progress	200	00					
progress	200	00					
progress	202	00					
progress	370 375	00					
progress	280	00					
progress	200	66					
progress	200	66					
progress	305	66					
progress	700	00					
co i fwo	400 20to	lopo	bog	thu	/ mo/		omoRo
8/130/1000	ante an b	ite	for	87 87	y . 11100 4192	ue L. D	eniobe
completer		000	in 1	04.4	+/0 7 mc	rat	e of
	1 40 1 To	tal	t i m/	$1 \cdot 1$, rai	e ur leted
Success.	1 10	lar		- · · L	23,	comp	leteu

estBuyModel@2acbc9b8 finishing with stats:

3680/s d Nov 17, 2014 10:06:41 AM

M²



new toos: 300 C State March 11 198



ef9053e4c5a713d0814a14947f216b4a1e7333bc,1214981,abcat0515039,"Logitech speakers","2011-08-11 09:03:58.306","2011-08-11 09:02:17.05 fa4bbed0c6a651c74db4eda0f523da8ff869e569,2573486,pcmcat218000050001,Asus,"2011-08-11 09:04:11.288","2011-08-11 09:00:34.853" e831860426098c8f9191318f71c56d05b7039d21,2496342,pcmcat209400050001,Xperia,"2011-08-11 09:04:11.796","2011-08-11 09:03:06.493" 29a921902ab04b3dc339f4b423cf824ea2869ec3,9097764,abcat0410018,10-20,"2011-08-11 09:04:14.571","2011-08-11 09:03:19.574" e6e83ccc269dc26b049f231139f52a1228b0b76c,1686194,abcat0703002,"Mass effect 2","2011-08-11 09:04:46.452","2011-08-11 09:04:19.712" 1bccd5893543a9324589a13b85286cd9af150121,2627475,cat02716,"Maybach music","2011-08-11 09:04:56.875","2011-08-11 09:04:41.844" d363ed695b34ef4e05695f9b81431b5e24757426,1232769,abcat0208011,"ipod speaker dock","2011-08-11 09:05:11.381","2011-08-11 09:02:01.21 f1c0aa8e2ecb1bb3e7cd87a8c0405557eb1fae10,9824298,pcmcat230800050019,"Turtle beach headset","2011-08-11 09:05:16.785","2011-08-11 09 c57fec4bb43bcfd2effbabe5a61c1f13cb52a06c,9835567,abcat0208031,Isimple,"2011-08-11 09:05:21.818","2011-08-11 09:05:02.714" a0ebda10ad523306a6830af4eef06482acf48cc2,9420361,pcmcat183800050007,"computer charger","2011-08-11 09:05:38.067","2011-08-11 09:04: 022969be550859b88fd48b4d3ce22ce443b4c434,1816383,pcmcat170900050018,Webcam,"2011-08-11 09:05:40.698","2011-08-11 09:02:39.729" c57fec4bb43bcfd2effbabe5a61c1f13cb52a06c,8760931,abcat0307018,Isimple,"2011-08-11 09:06:35.152","2011-08-11 09:05:02.714" 097c9ed5549d5ba, 168049, cath2719, jay-z, "2011-18-11 09:06:39.825", "2011-0_11 09:06:23.153" 9a113240869815c067a369e

d8c14c4a46793addd1d0f11e326bc2fce89eefca,3045049,pcmcat212600050008,"hp_dual_core","2011-08-11_09:07:09.067","2011-08-11_09:04:34.5

4113ac371d0c1964156065da7d



b9df0b9099c87224c3ea60eed48d47aeb53d85e4,1051329,abcat0715007,"Ps3 move","2011-08-11 09:07:51.897","2011-08-11 09:06:37.895" 723bea69e272a99ec79b818b6f983b138d1a8ab5,9225377,abcat0201011,"iPod nano","2011-08-11 09:07:57.449","2011-08-11 09:06:58.583" 9f7fbe852db0d845d2bf43472c4f4b54490c861b,8280834,abcat0107004,"hdtv antenna","2011-08-11 09:08:11.619","2011-08-11 09:06:49.658" 8e7d1caf3c3fd56bee3697f63188be61c630048a,3048064,pcmcat247400050000,"Lap tops","2011-08-11 09:08:26.366","2011-08-11 09:07:51.508" 3920405da68ea703996bc425893e3f9d2aa41648,2128267,abcat0801003,"virgin mobile","2011-08-11 09:08:36.533","2011-08-11 09:07:11.16" b3d51f75dc3e595cbf5871bae9c8d5b0656aa51c,2408109,abcat0101001,"65 inch tv","2011-08-11 09:08:48.206","2011-08-11 09:07:24.274" 3658871850982b856204b9a338127ef8a110b0dc,9952217,pcmcat151600050006,"Fender starcaster","2011-08-11 09:09:00.872","2011-08-11 09:08 c9544776252f9334afea7416a0df54df4fe08eed,9492426,pcmcat143000050007,"beats by dre","2011-08-11 09:09:01.29","2011-08-11 09:08:12.56 90b1b7d9a5c27a98428901735f255a12ec4afb71,2642464,abcat0401004,cameras,"2011-08-11 09:09:03.367","2011-08-11 09:04:46.199" f8cbee1885e089c869e78d6d0e6c105aafc42bac,7898082,abcat0503003,"Wireless n","2011-08-11 09:09:14.413","2011-08-11 09:07:37.624" bd0add05492ec9301c902345c57bed20d8c303bc,2715258,pcmcat247400050000,intel,"2011-08-11 09:09:16.892","2011-08-11 09:07:08.29" c5c73ded3bbe43f9baddd220d0461e89a042fa6e,2822239,abcat0307015,auxiliary,"2011-08-11 09:09:18.47","2011-08-11 09:06:38.533"



11 09:05:56.298" 3. 54","2011-08-11 09:05 08-11 09:06:49.087"

4cde5c0acecab3dfe51bb260e3490bfc882b6f41,2408109,abcat0101001,"Panasonic gt30","2011-08-11 09:07:13.387","2011-08-11 09:06:53.312" 953dc1bb1c9d5145a50ffae5a676de605d8005e2,1167091,pcmcat143000050011,"Skullcandy low","2011-08-11 09:07:22.77","2011-08-11 09:06:32. 7da552b3,1158093,abcat051102 ... "2011-08- ... 09:07:25.379","2011-08-11 09:05:48.273"

-08-11 09:06:49.087"



data warehousing









$\log 4$ transform 4 use

Data Architecture

dimensional modeling

relational modeling

Warehouse Design

normalization

denormalization

star schema

slowly changing dimensions

slowly changing dimensions

type 1: overwrite the old data

type 2: multiple rows with versioning

type 3: extra columns for older versions

slowly changing dimensions

event history





Production





Production Development

trait EventHistory { def publishEvent(e: Event)

def getEvents (startTime: Date, endTime: Date,

- eventFilter: EventFilter,
- eventHandler: EventHandler

event history design

Data Architecture

log a lot

ok to denormalize

think about the types
- Make sure that events are simple facts
- Files are ok for event history, don't really need a database
- Use an object hierarchy to model events in code
- Use online features that are efficient to update incrementally
- Write efficient implementations before than scaling out
- Functional style makes it easier
- Encourage reactive processing





- Matters more than transformations, more than algorithms
- Data that doesn't make sense often indicates an application bug
- Do assertions, e.g., make sure things aren't happening out of order

Data Quality



- All data in form of events no exceptions!
- Same feature code in production and development
- Emphasis on creative feature engineering
- Quick cycles between ideas and production

Try the Antelope Demo: <u>https://github.com/ifwe/antelope/blob/master/doc/demo.md</u>

github.com/ifwe/antelope @jssmith

