SOLUTION TRACK

Finding the Needle in a Big Data Haystack

@EvaAndreasson, Innovator & Problem Solver Cloudera



Agenda

- Problem (Solving)
- Apache Solr + Apache Hadoop et al
- Real-world examples
- Q&A



Problem Solving





Information Driven Problem Solving

- Ask a Question
- Find All Relevant Data to Serve the Question
- Process the Data to Answer the Question





Information Driven Businesses



Problem: Finding the Data (Needle) Across (Hay) Silos

Thousands of employees & lots of data Difficult to access

Heterogeneous legacy IT Infrastructure Difficult to manage Hard to scale



©2014 Cloudera, Inc. All Rights Reserved. Reproduction or redistribution without written permission is prohibited.

cloudera

Solution: The Enterprise Data Hub (EDH)



©2014 Cloudera, Inc. All Rights Reserved. Reproduction or redistribution without written permission is prohibited.



Apache Hadoop et al at the Core

- Open source
 - white box, best innovation at all times
- Flexible
- Scalable
 - ingest, storage, processing
- Cost efficient



But an EDH also Needs...

- Security & audit
- Manageability, Visibility and Resource Control
- Open architecture
- Multi-workload support and optimization



CLOURERA ERH



Problem solved! We can go home...?



Problem: Finding the Needle in a Big Data Haystack?





New Audiences, New Challenges

- Non-technical staff needs access to data
 - Same data used in bigger processes
 - Speed up manual introspection
- Technical staff needs to view / explore data
 - View interim results
 - Drill down into mission critical data
 - Explore data to design models
- Cross-workload needs
 - Combine structured and unstructured data



Solution: Everyone Knows Search!





Cloudera Search: How We Integrated Solr with Hadoop et al

Cloudera Search

Interactive search for Hadoop

- Full-text and faceted navigation
- Batch, near real-time, and on-demand indexing

Apache Solr integrated with CDH

- Established, mature search with vibrant community
- Separate runtime like MapReduce, Impala
- Incorporated as part of the Hadoop ecosystem

Open Source

- 100% Apache, 100% Solr
- Standard Solr APIs





Scalable and Robust Index Storage



Solr and HDFS

- Scalable, cost-efficient index storage
- Higher availability
- Search and process data in one platform



Near Real Time Indexing at Ingest



Solr and Flume

- Data ingest at scale
- Flexible extraction and mapping
- Indexing at data ingest
- Document-level ACL



Scalable Batch Indexing



Solr and MapReduce

- Flexible, scalable batch indexing
- GOLIVE: Start serving new indices with no downtime
- On-demand indexing, costefficient re-indexing



Scalable Batch Indexing





Secondary Indexes, in Real-Time





Secondary Indexes, or in Batch





Streamlined Extraction and Mapping



Morphlines

- Simple and flexible data transformation
- Reusable across multiple index (and other) workloads



Security

Cloudera Search + Sentry

- Cluster level access control
- Index level access control



Simple, Customizable Search UI

Hue 💱 😚 🗖	🛋 🔯 🤇	0 🌿 💷 9 3 0			
Search in tweets apple		O Go			
USER FOLLOWERS COUNT	Showing	1 - 15 of 508 results			
0 (35) 20 (18) 30 (12) 40 (12) 50 (15) 60 (11) 70 (6)		Frankle Sinclair Urban Splash property. A bit like Apple: overpriced and fab design. Except with Appl ← Reply 12 Retweet Kaylee Navarra Also I find it funny how, if you don't have a warranty on your apple product, you have			
80 (11) 90 (13) LOCATION seattle (4) usa (4) venezuela (4)		 ← Reply 12 Retweet stix aluco RT @HiginiaRoig: Mc'Donalds substitueix la Catalonia. Mango es menja la llibreria #0 El mercat e ← Reply 12 Retweet 			
united states (3) canada (2) cape town, south africa (2) chicago, il (2) ohio (2)	and the	Mohamad kaaki RT @WAnaShajra: Wa2ta tes2al benet "are you virgin ?" Betred "No im apple store" a3mlik 3asir frais s Reply 13 Retweet			
san jose, ca (2) RETWEETS 0 (508)	R.	Bryan Thomas RT @MacRumors: Apple Releases Economic Impact Report on 'Spaceship Campus			
		Beth Green apples are so sour ewww 😺			
	T	Leonardo reyes RT @EspecJarocho: #Descubre IRADIO SERÁ EL NOMBRE DEL NUEVO SERVICIO OFRECERÁN http://t.co/WUguRpA1qe http://t.co/			
		bartiriamarinho RT @folha_com: Android cresce e está perto de alcançar número de aplicativos bai			

Hue

- Simple UI
- Navigated, faceted drill down
- Customizable display
- Full text search, standard Solr API and query language



Simplified Management

cloudera manager										
Home	Services -	Hosts	Activities +	Diagno	ose 🕶	Audits	Cha			
Services »	Service SOI P-1	22								
, SOL	.R-1									
🔒 Status		O Comma	nds 🛛 Confi	guration •	• 👁 Aud	lits 🔁 V	/eb UI			
Role Ins	Role Instances									
Actions for Selected - Add										
🔒 Nam	е	Host		R	ole Group		Rack			
Any	/ Role Type	 Any H 	ost				Any			
solr_	server (sys-viks4	I-1) sys-vik	s4-1.ent.cloude	ra.com So	olr Server ((Default)	/defau			
⊜ <u>solr</u> _	server (sys-viks4	<u>1-2) sys-vik</u>	s4-2.ent.cloude	ra.com So	olr Server ((Default)	/defaı			
⊜ <u>solr</u>	server (sys-viks4	I <u>-3)</u> <u>sys-vik</u>	s4-3.ent.cloude	ra.com So	olr Server ((Default)	/defau			
i solr_	server (sys-viks4	<u>1-4) sys-vik</u>	s4-4.ent.cloude	r <u>a.com</u> So	olr Server ((Default)	/defai			

Cloudera Manager

- Install, configure, deploy SolrCloud on the cluster
- Centralized management and monitoring – cross workloads
- Unified resource management and control



Architecture Overview



Some Real-World Examples

- Image processing and data correlation
 - Quick result checks on new algorithms
 - Correlate image and device logs or external data
 - Image exploration as a service
 - Expedite data modeling processes
- Free-text form matching
 - Claims report correlation, to gather data likely to be similar
 - Serve 360-client or patient records in a speedier way
 - Fraud / pattern extraction
- Log management
 - Real-time drill down
 - Long term trending and capacity planning
 - Anomaly detection over larger sets of data



The EDH - Information-Driven Problem Solving Made Easy!





Integration is Key

- Eliminate data moves or copies, break silos
- Create a truly active archive
- Serve non-technical audiences on the same platform as where advanced analytics workloads run
- Analyze and combine structured and non-structured data
- Expedite exploration of various data types
- Find data and do something with it where it is stored
- Future proof your data management system



Learn More

- Cloudera.com
 - Read our blog
 - Take our online training (or get Cloudera certified)
 - Download whitepapers
 - View webinars
 - Talk to our customers
- Follow/contact me
 - @EvaAndreasson





Please evaluate my talk via the mobile app!





COUDERS® Ask Bigger Questions