Introduction to Big Data with Apache Spark

BerkeleyX

This Lecture

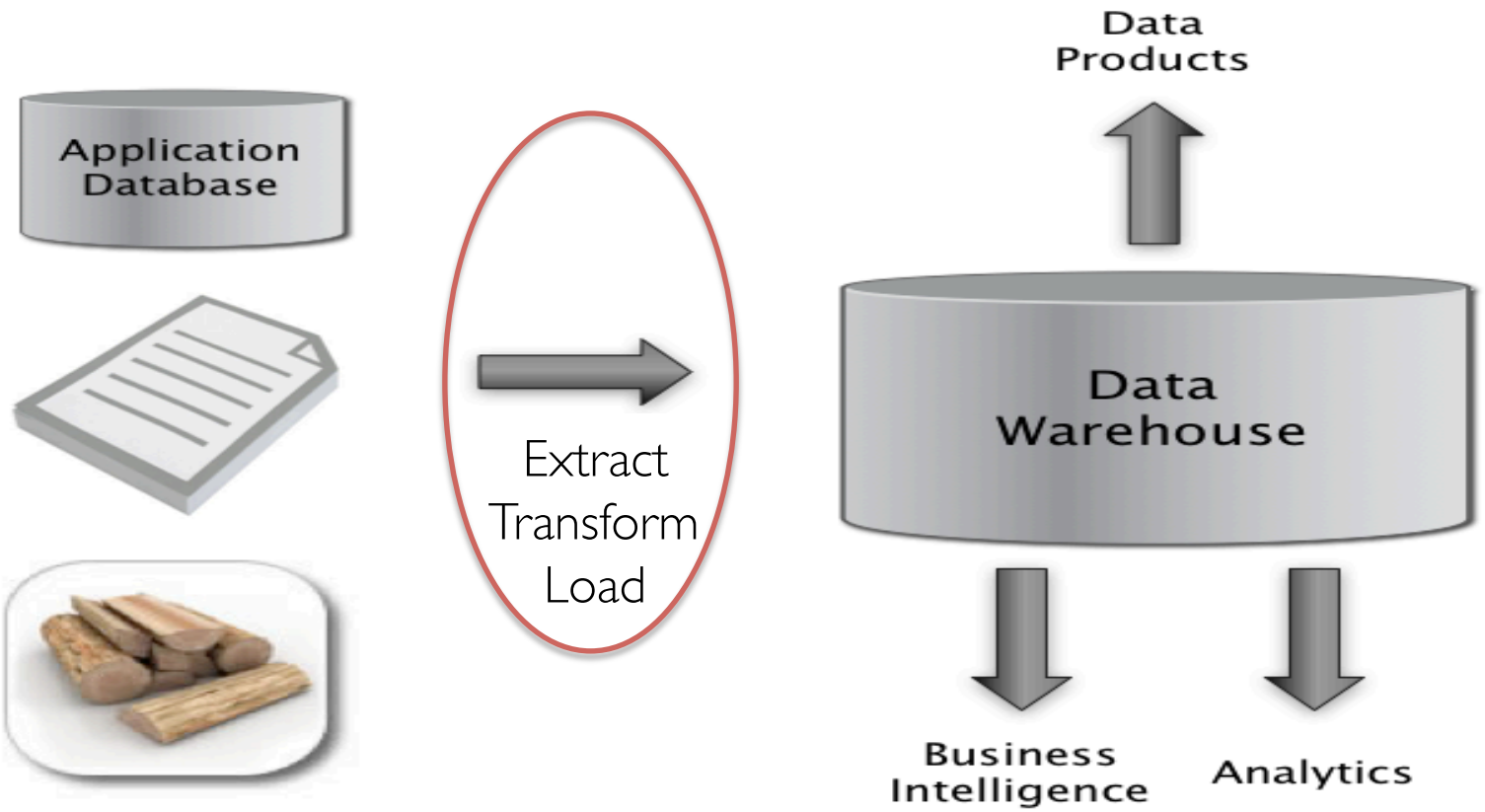
The Structure Spectrum

Files: Formats and Performance

Tabular Data: Examples, Challenges, pySpark DataFrames

Log Files

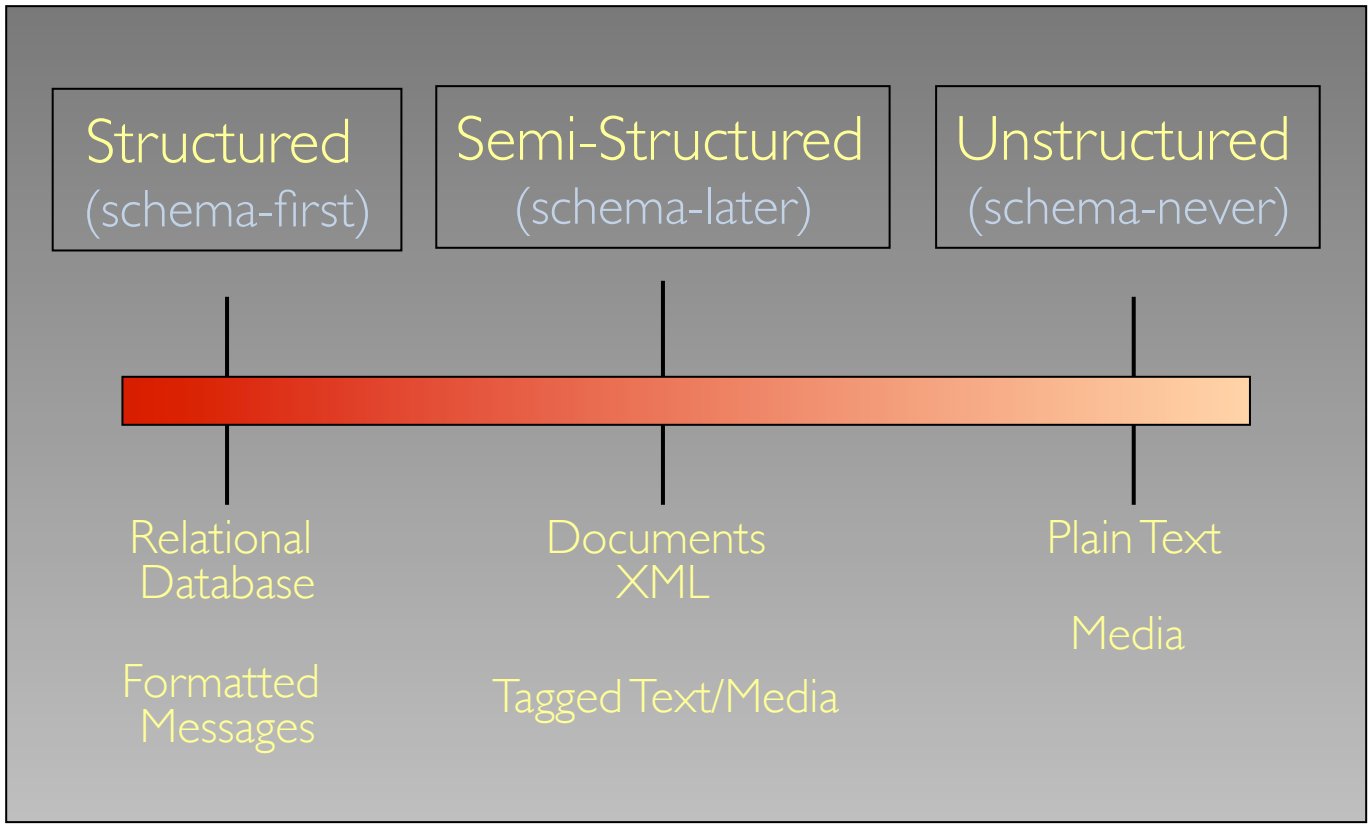
Review: The Big Picture



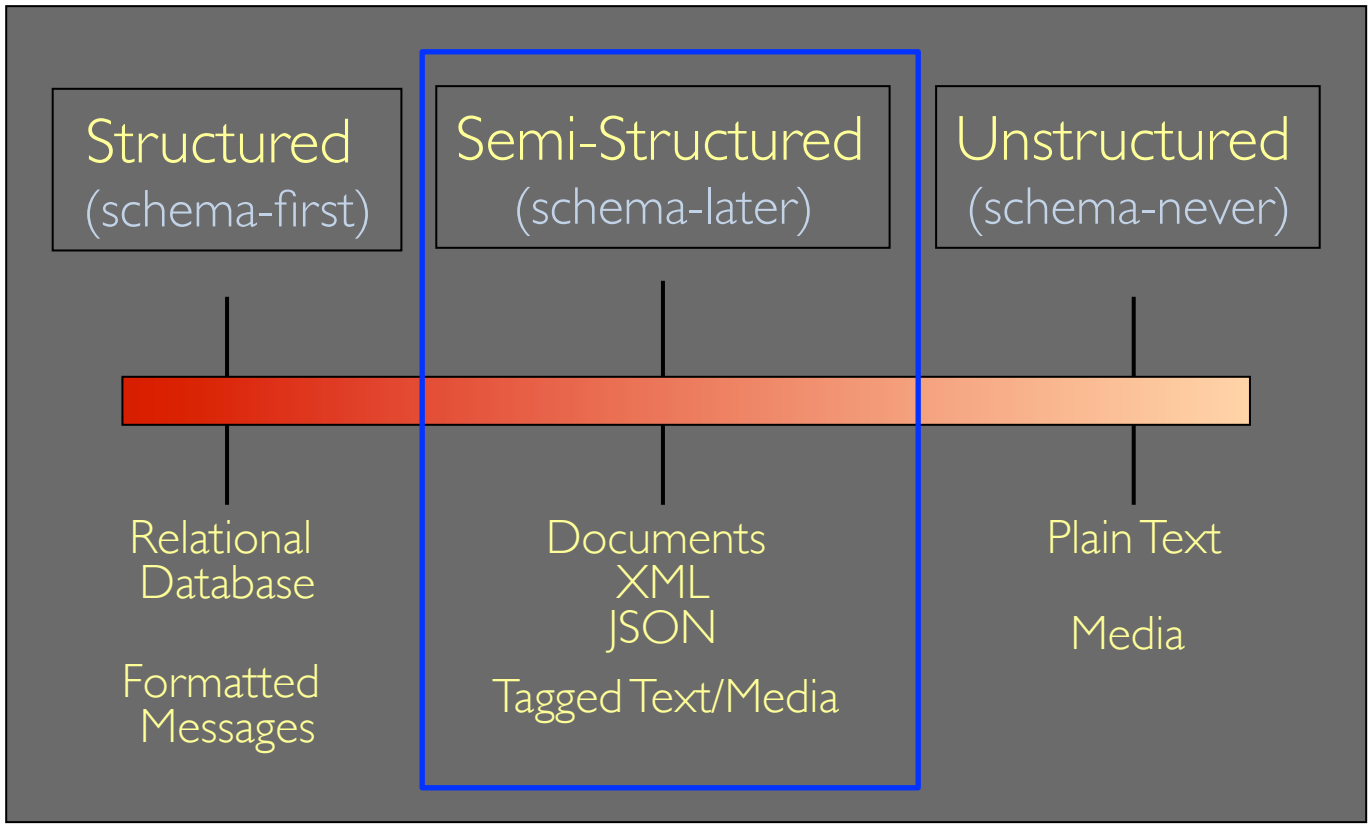
Key Data Management Concepts

- A *data model* is a collection of concepts for describing data
- A *schema* is a description of a particular collection of data, using a given data model

The Structure Spectrum



The Structure Spectrum



This lecture

Files

- What is a file?
  - » A *file* is a named sequence of *bytes*
    - Typically stored as a collection of pages (or blocks)
  - » A *filesystem* is a collection of files organized within an hierarchical namespace
    - Responsible for laying out those bytes on physical media
    - Stores file metadata
    - Provides an API for interaction with files
  - » Standard operations
    - `open()/close()`
    - `seek()`
    - `read()/write()`

http://www.clipshrine.com/Document-Icon-8906-medium.html

Files: Hierarchical Namespace

- On Linux, / is the root of a filesystem
- On Windows, \ is the root of a filesystem
- Files and and directories have associated permissions
- Files are not always arranged in a hierarchically
  - » Content-addressable storage (CAS)
  - » Often used for large multimedia collections

Considerations for a File Format

- Data model: tabular, hierarchical, array
- Physical layout
- Field units and validation
- Metadata: header, side file, specification, other?
- Plain text (ASCII, UTF-8, other) or binary
- Delimiters and escaping
- Compression, encryption, checksums?
- Schema evolution

Semi-Structured Tabular Data

- *One of the most common data formats*
- A *table* is a collection of *rows* and *columns*
- Each row has an *index* and each column has a *name*
- A *cell* is specified by an (index, name) *pair*
- A cell may or may not have a *value*
- A cell's *type* is inferred from its value

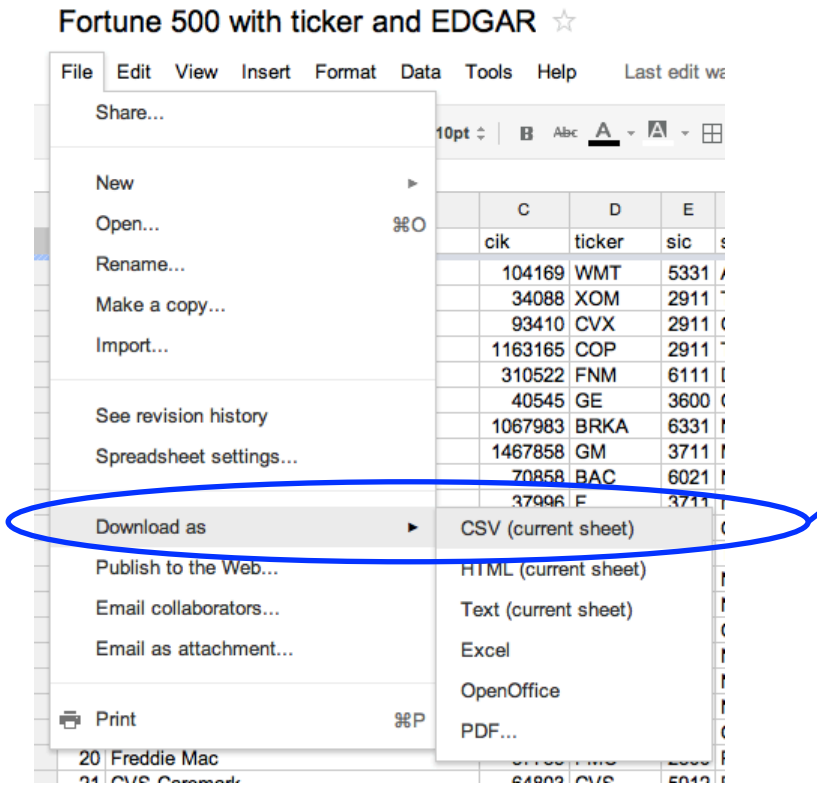
Tabular Data Example

- Fortune 500 companies
  - » Top 500 US closely held and public corporations by gross revenue

	A	B	C	D	E	F	G	H	I
1	rank	company	cik	ticker	sic	state_location	state_of_incorporation	revenues	profits
2	1	Wal-Mart Stores	104169	WMT	5331	AR	DE	421849	16389
3	2	Exxon Mobil	34088	XOM	2911	TX	NJ	354674	30460
4	3	Chevron	93410	CVX	2911	CA	DE	196337	19024
5	4	ConocoPhillips	1163165	COP	2911	TX	DE	184966	11358
6	5	Fannie Mae	310522	FNM	6111	DC	DC	153825	-14014
7	6	General Electric	40545	GE	3600	CT	NY	151628	11644
8	7	Berkshire Hathaway	1067983	BRKA	6331	NE	DE	136185	12967
9	8	General Motors	1467858	GM	3711	MI	MI	135592	6172
10	9	Bank of America Corp.	70858	BAC	6021	NC	DE	134194	-2238
11	10	Ford Motor	37996	F	3711	MI	DE	128954	6561
12	11	Hewlett-Packard	47217	HPQ	3570	CA	DE	126033	8761
13	12	AT&T	732717	T	4813	TX	DE	124629	19864
14	13	J.P. Morgan Chase & Co.	19617	JPM	6021	NY	DE	115475	17370
15	14	Citigroup	831001	C	6021	NY	DE	111055	10602
16	15	McKesson	927653	MCK	5122	CA	DE	108702	1263
17	16	Verizon Communications	732712	VZ	4813	NY	DE	106565	2549
18	17	American International Group	5272	AIG	6331	NY	DE	104417	7786
19	18	International Business Machines	51143	IBM	3570	NY	NY	99870	14833
20	19	Cardinal Health	721371	CAH	5122	OH	OH	98601.9	642.2
21	20	Freddie Mac	37785	FMC	2800	PA	DE	98368	-14025

http://fortune.com/fortune500/

Exporting Tabular Data



- US Fortune 500

Export as Comma Separated Values

http://fortune.com/fortune500/

Tabular Data (CSV File)

Fortune 500 with ticker and EDGAR – Plus Ticker and EDGAR.txt

```
rank,company,cik,ticker,sic,state_location,state_of_incorporation,revenues,profits
1,Wal-Mart Stores,104169,WMT,5331,AR,DE,421849,16389
2,Exxon Mobil,34088,XOM,2911,TX,NJ,354674,30460
3,Chevron,93410,CVX,2911,CA,DE,196337,19024
4,ConocoPhillips,1163165,COP,2911,TX,DE,184966,11358
5,Fannie Mae,310522,FNM,6111,DC,DC,153825,-14014
6,General Electric,40545,GE,3600,CT,NY,151628,11644
7,Berkshire Hathaway,1067983,BRKA,6331,NE,DE,136185,12967
8,General Motors,1467858,GM,3711,MI,MI,135592,6172
9,Bank of America Corp.,70858,BAC,6021,NC,DE,134194,-2238
10,Ford Motor,37996,F,3711,MI,DE,128954,6561
11,Hewlett-Packard,47217,HPQ,3570,CA,DE,126033,8761
12,AT&T,732717,T,4813,TX,DE,124629,19864
13,J.P. Morgan Chase & Co.,19617,JPM,6021,NY,DE,115475,17370
14,Citigroup,831001,C,6021,NY,DE,111055,10602
15,McKesson,927653,MCK,5122,CA,DE,108702,1263
16,Verizon Communications,732712,VZ,4813,NY,DE,106565,2549
17,American International Group,5272,AIG,6331,NY,DE,104417,7786
18,International Business Machines,51143,IBM,3570,NY,NY,99870,14833
19,Cardinal Health,721371,CAH,5122,OH,OH,98601.9,642.2
20,Freddie Mac,37785,FMC,2800,PA,DE,98368,-14025
21,CVS Caremark,64803,CVS,5912,RI,DE,96413,3427
22,UnitedHealth Group,731766,UNH,6324,MN,MN,94155,4634
23,Wells Fargo,72971,WFC,6021,CA,DE,93249,12362
24,Valero Energy,1035002,VLO,2911,TX,DE,86034,324
25,Kroger,56873,KR,5411,OH,OH,82189.4,1116.3
26,Procter & Gamble,80424,PG,2840,OH,OH,79689,12736
27,AmerisourceBergen,1140859,ABC,5122,PA,DE,77954,636.7
28,Costco Wholesale,909832,COST,5331,WA,WA,77946,1303
29,Marathon Oil,101778,MRO,2911,TX,DE,68413,2568
30,Home Depot,354950,HD,5211,GA,DE,67997,3338
```

http://fortune.com/fortune500/

- US Fortune 500

Comma Separated Values Format:
Field1, Field2, Field3, Field4, …
Value1_1, Value1_2, Value1_3, Value1_4,…
Value2_1, Value2_2, Value2_3, Value2_4, …
…

Protein Data Bank

```
HEADER    APOPTOSIS                               23-DEC-12   3J2T
TITLE     AN IMPROVED MODEL OF THE HUMAN APOPTOSOME
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: APOPTOTIC PROTEASE-ACTIVATING FACTOR 1;
COMPND   3 CHAIN: A, B, C, D, E, F, G;
COMPND   4 SYNONYM: APAF-1;
COMPND   5 ENGINEERED: YES;
COMPND   6 MOL_ID: 2;
COMPND   7 MOLECULE: CYTOCHROME C;
COMPND   8 CHAIN: H, I, J, K, L, M, N
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE   3 ORGANISM_COMMON: HUMAN;
SOURCE   4 ORGANISM_TAXID: 9606;
SOURCE   5 GENE: APAF-1, APAF1, KIAA0413;
SOURCE   6 EXPRESSION_SYSTEM: SPODOPTERA FRUGIPERDA;
SOURCE   7 EXPRESSION_SYSTEM_COMMON: FALL ARMYWORM;
KEYWDS    APOPTOSIS PROTEASE ACTIVATING FACTOR-1, APAF-1, CYTOCHROME C,
KEYWDS   2 APOPTOSIS
EXPDTA    ELECTRON MICROSCOPY
AUTHOR    S.YUAN,M.TOPF,C.W.AKEY
REVDAT   2   17-APR-13 3J2T    1       JRNL
REVDAT   1   10-APR-13 3J2T    0
```

PDB Format:
Field #, Values
Field #, Values
Field #, Values
...

http://www.rcsb.org/pdb/files/3J2T.pdb

Tabular Data

Several Challenges

- Format not well-defined (may be missing data values)
- Types may be incorrectly inferred (“2” versus “2.0”)
- No support for versioning of format
- …

Tabular Data from Multiple Sources

Several Challenges

- May be missing fields (not every source provides same data)
- Inconsistent data types (one file has $ values another has £)
- Inconsistent values for same entity (Wal-Mart versus WalMart)
- …

Tabular Data from Sensors

Several Challenges

- May be missing fields (a given sensor may not produce all types)
- Sensor may be damaged (permanently or intermittently)
- Timestamps may not be accurate
- Other metadata (sensor location, ID) may have errors
- Sensor may go offline for a while
- …

pandas: Python Data Analysis Library

- Open source data analysis and modeling library
  - » An alternative to using R

- pandas DataFrame: a table with named columns
  - » The most commonly used pandas object
  - » Represented as a Python Dict (column_name ➔ Series)
  - » Each pandas Series object represents a column
    - 1-D labeled array capable of holding any data type
  - » R has a similar data frame type

Semi-Structured Data in pySpark

- DataFrames introduced in Spark 1.3 as extension to RDDs
- *Distributed* collection of data organized into named columns
  - » Equivalent to Pandas and R DataFrame, but distributed
- Types of columns inferred from values

pySpark and pandas DataFrames

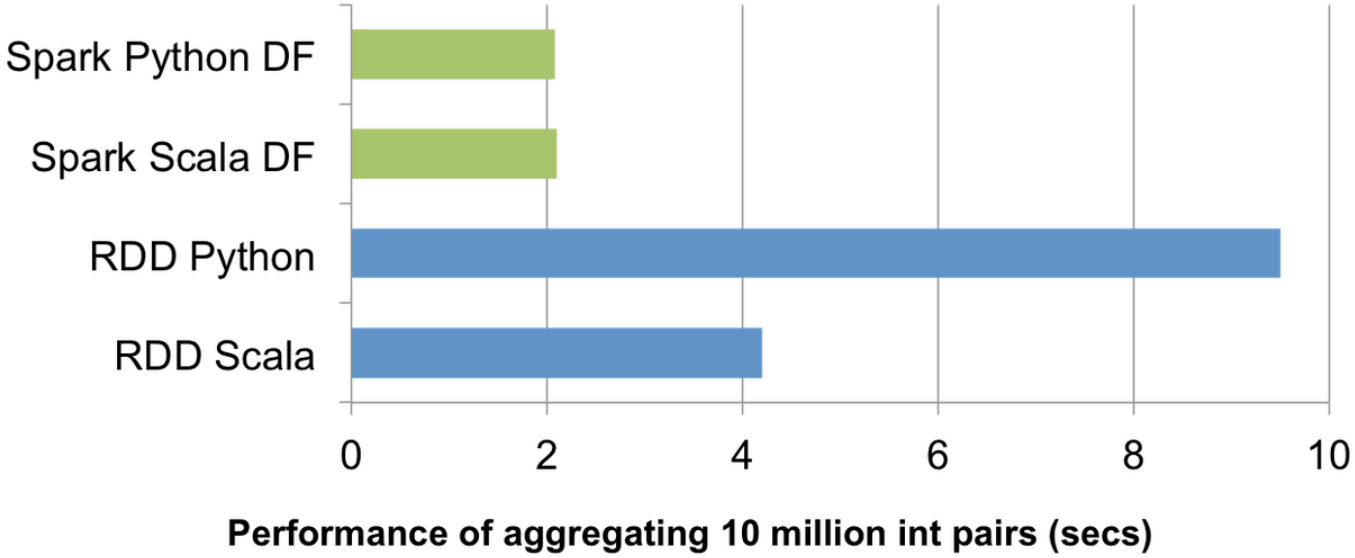
- Easy to convert between Pandas and pySpark
  - » *Note: pandas DataFrame must fit in driver*

```
# Convert Spark DataFrame to Pandas
pandas_df = spark_df.toPandas()

# Create a Spark DataFrame from Pandas
spark_df = context.createDataFrame(pandas_df)
```

pySpark DataFrame Performance

- Almost 5x pySpark performance on a single machine



https://databricks.com/blog/2015/02/17/introducing-dataframes-in-spark-for-large-scale-data-science.html

Semi-Structured Log Files

- Created by **`printf`** statements in server processes:
  - » Web, database, network file servers, operating system components
- Human-readable text format files
  - » Very rarely actually read by a human
  - » Can store/archive in binary or compressed format
- Format published or "defined" by code
  - » Can be very difficult to parse

Recall: Apache Web Server Log

```
uplherc.upl.com - - [01/Aug/1995:00:00:07 -0400] "GET / HTTP/1.0" 304 0
uplherc.upl.com - - [01/Aug/1995:00:00:08 -0400] "GET /images/ksclogo-medium.gif
HTTP/1.0" 304 0
uplherc.upl.com - - [01/Aug/1995:00:00:08 -0400] "GET /images/MOSAIC-logosmall.gif
HTTP/1.0" 304 0
uplherc.upl.com - - [01/Aug/1995:00:00:08 -0400] "GET /images/USA-logosmall.gif HTTP/
1.0" 304 0
ix-esc-ca2-07.ix.netcom.com - - [01/Aug/1995:00:00:09 -0400] "GET /images/launch-
logo.gif HTTP/1.0" 200 1713
uplherc.upl.com - - [01/Aug/1995:00:00:10 -0400] "GET /images/WORLD-logosmall.gif
HTTP/1.0" 304 0
slppp6.intermind.net - - [01/Aug/1995:00:00:10 -0400] "GET /history/skylab/
skylab.html HTTP/1.0" 200 1687
piweba4y.prodigy.com - - [01/Aug/1995:00:00:10 -0400] "GET /images/launchmedium.gif
HTTP/1.0" 200 11853
tampico.usc.edu - - [14/Aug/1995:22:57:13 -0400] "GET /welcome.html HTTP/1.0" 200 790
```

Apache Web Server Log Format

- Apache Common Log Format specifies log file format
- Example line from log file:
  - `127.0.0.1 - - [01/Aug/1995:00:00:01 -0400] "GET /images/launch-logo.gif HTTP/1.0" 200 1839`
- Components:
  - **`127.0.0.1`** *Client IP address*

Apache Web Server Log Format

- Apache Common Log Format specifies log file format
- Example line from log file:
  - `127.0.0.1 - - [01/Aug/1995:00:00:01 -0400] "GET /images/launch-logo.gif HTTP/1.0" 200 1839`
- Components:
  - **-** *User identity from remote machine* (hyphen means not available)
  - **-** *User identity from local logon* (hyphen means not available)

Apache Web Server Log Format

- Apache Common Log Format specifies log file format
- Example line from log file:
  - `127.0.0.1 - - [01/Aug/1995:00:00:01 -0400] "GET /images/launch-logo.gif HTTP/1.0" 200 1839`
- Components:
  - **`[01/Aug/1995:00:00:01 -0400]`**

    Request time

Apache Web Server Log Format

- Apache Common Log Format specifies log file format
- Example line from log file:
  - `127.0.0.1 - - [01/Aug/1995:00:00:01 -0400] "GET /images/launch-logo.gif HTTP/1.0" 200 1839`
- Components:
  - **`"GET /images/launch-logo.gif HTTP/1.0"`**

    Client request
    - Request method (e.g., GET, POST, etc.)
    - Endpoint (a Uniform Resource Identifier)
    - Client protocol version

Apache Web Server Log Format

- Apache Common Log Format specifies log file format
- Example line from log file:
  - `127.0.0.1 - - [01/Aug/1995:00:00:01 -0400] "GET /images/launch-logo.gif HTTP/1.0" 200 1839`
- Components:
  - **`200`** *Status code the server sent back to the client*
    - OK response (2xx), others: 3xx, 4xx, 5xx
  - **`1839`** *Size of the object returned to client*
    "-" if no content returned, ***or sometimes 0***

Lab: Explore Web Server Access Log

- NASA HTTP server access log
  - » http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html
- Log covers 21 days (1 Aug, 3 Aug – 22 Aug 1995)
- Log includes 1,043,177 requests
- Log partially cleaned for you
  - » Removed some very hard to parse requests

Some Log Analysis Questions

- Overall:
  - » What are the statistics for content being returned? Sizes, statuses?
  - » What are the types of return codes?
  - » How many 404 (page not found) errors are there?

- Temporal:
  - » How many unique hosts per day?
  - » How many requests per day?
  - » On average, how many requests per host?
  - » How many 404 errors per day?

Recall: Machine System Log File

```
dhcp-47-129:CS100_1> syslog -w 10
Feb  3 15:18:11 dhcp-47-129 Evernote[1140] <Warning>: -[EDAMAccounting
read:]: unexpected field ID 23 with type 8.  Skipping.
Feb  3 15:18:11 dhcp-47-129 Evernote[1140] <Warning>: -[EDAMUser read:]:
unexpected field ID 17 with type 12.  Skipping.
Feb  3 15:18:11 dhcp-47-129 Evernote[1140] <Warning>: -
[EDAMAuthenticationResult read:]: unexpected field ID 6 with type 11.
Skipping.
Feb  3 15:18:11 dhcp-47-129 Evernote[1140] <Warning>: -
[EDAMAuthenticationResult read:]: unexpected field ID 7 with type 11.
Skipping.
Feb  3 15:18:11 dhcp-47-129 Evernote[1140] <Warning>: -[EDAMAccounting
read:]: unexpected field ID 19 with type 8.  Skipping.
Feb  3 15:18:11 dhcp-47-129 Evernote[1140] <Warning>: -[EDAMAccounting
read:]: unexpected field ID 23 with type 8.  Skipping.
Feb  3 15:18:11 dhcp-47-129 Evernote[1140] <Warning>: -[EDAMUser read:]:
unexpected field ID 17 with type 12.  Skipping.
Feb  3 15:18:11 dhcp-47-129 Evernote[1140] <Warning>: -[EDAMSyncState read:]:
unexpected field ID 5 with type 10.  Skipping.
Feb  3 15:18:49 dhcp-47-129 com.apple.mtmd[47] <Notice>: low priority

thinning needed for volume Macintosh HD (/) with 18.9 <= 20.0 pct free space
```

splunk>

Data Mining ("Splunking")

- Collect files from many machines
  - » Application and system event logs
- Check for unusual events:
  - » Disk errors
  - » Network congestion
  - » Security attacks
- Monitor resources
  - » Network, memory, disk, CPU, application queues
- Visualize with a dashboard
  - » Splunk Free License

http://www.splunk.com

Considerations for a File Format

- Data model: tabular, hierarchical, array
- Physical layout
- Field units and validation
- Metadata: header, side file, specification, other?
- Plain text (ASCII, UTF-8, other) or binary
- Delimiters and escaping
- Compression, encryption, checksums?
- Schema evolution

Performance!

File Performance Considerations

- Read versus write performance
- Plain text versus binary format
- Environment: Pandas (Python) versus Scala/Java
- Uncompressed versus compressed

File Performance

626 MB text file
787 MB binary file

	Read Time (Text)	Write Time (Text)	Read Time (Binary)	Write Time (Binary)
Pandas (Python)	36 secs	45 secs	**	**
Scala/Java	18 secs	21 secs	1-6* secs	1-6* secs

Read-Write Times Comparable

** Pandas doesn't have a binary file I/O library
(Python performance depends on library you use)

* 6 seconds is the time for sustained read/write
(often faster due to system caching)

File Performance

626 MB text file
787 MB binary file

	Read Time (Text)	Write Time (Text)	Read Time (Binary)	Write Time (Binary)
Pandas (Python)	36 secs	45 secs	**	**
Scala/Java	18 secs	21 secs	1-6* secs	1-6* secs

Binary I/O much faster than text

** Pandas doesn't have a binary file I/O library
(Python performance depends on library you use)

* 6 seconds is the time for sustained read/write
(often faster due to system caching)

File Performance - Compression

Binary File	Read Time	Write Time	File Size
Gzip level 6 (Java default)	4 secs	75 secs	286 MB
Gzip level 3	4 secs	20 secs	
Gzip level 1	4 secs	14 secs	328 MB
LZ4 fast	2 secs	4 secs	423 MB
Raw binary file	1-6 secs	1-6 secs	787 MB

Text File	Read Time	Write Time	File Size
Gzip level 6 (default)	26 secs	98 secs	243 MB
Gzip level 3	25 secs	46 secs	259 MB
Gzip level 1	25 secs	33 secs	281 MB
LZ4 fast	22 secs	24 secs	423 MB
Raw text file	18 secs	21 secs	626 MB

Scala/Java language

Write times much larger than read

File Performance - Compression

Binary File	Read Time	Write Time	File Size
Gzip level 6 (Java default)	4 secs	75 secs	286 MB
Gzip level 3	4 secs	20 secs	
Gzip level 1	4 secs	14 secs	328 MB
LZ4 fast	2 secs	4 secs	423 MB
Raw binary file	1-6 secs	1-6 secs	787 MB

Text File	Read Time	Write Time	File Size
Gzip level 6 (default)	26 secs	98 secs	243 MB
Gzip level 3	25 secs	46 secs	259 MB
Gzip level 1	25 secs	33 secs	281 MB
LZ4 fast	22 secs	24 secs	423 MB
Raw text file	18 secs	21 secs	626 MB

Scala/Java language

Large range of compression times

File Performance - Compression

Binary File	Read Time	Write Time	File Size
Gzip level 6 (Java default)	4 secs	75 secs	286 MB
Gzip level 3	4 secs	20 secs	
Gzip level 1	4 secs	14 secs	328 MB
LZ4 fast	2 secs	4 secs	423 MB
Raw binary file	1-6 secs	1-6 secs	787 MB

Text File	Read Time	Write Time	File Size
Gzip level 6 (default)	26 secs	98 secs	243 MB
Gzip level 3	25 secs	46 secs	259 MB
Gzip level 1	25 secs	33 secs	281 MB
LZ4 fast	22 secs	24 secs	423 MB
Raw text file	18 secs	21 secs	626 MB

Scala/Java language

Large range of compression times

File Performance - Compression

Binary File	Read Time	Write Time	File Size
Gzip level 6 (Java default)	4 secs	75 secs	286 MB
Gzip level 3	4 secs	20 secs	313 MB
Gzip level 1	4 secs	14 secs	328 MB
LZ4 fast	2 secs	4 secs	423 MB
Raw binary file	1-6 secs	1-6 secs	787 MB

Text File	Read Time	Write Time	File Size
Gzip level 6 (default)	26 secs	98 secs	243 MB
Gzip level 3	25 secs	46 secs	259 MB
Gzip level 1	25 secs	33 secs	281 MB
LZ4 fast	22 secs	24 secs	423 MB
Raw text file	18 secs	21 secs	626 MB

Scala/Java language

Small range (15%) of compressed file sizes

File Performance - Compression

Binary File	Read Time	Write Time	File Size
Gzip level 6 (Java default)	4 secs	75 secs	286 MB
Gzip level 3	4 secs	20 secs	313 MB
Gzip level 1	4 secs	14 secs	328 MB
LZ4 fast	2 secs	4 secs	423 MB
Raw binary file	1-6 secs	1-6 secs	787 MB

Text File	Read Time	Write Time	File Size
Gzip level 6 (default)	26 secs	98 secs	243 MB
Gzip level 3	25 secs	46 secs	259 MB
Gzip level 1	25 secs	33 secs	281 MB
LZ4 fast	22 secs	24 secs	423 MB
Raw text file	18 secs	21 secs	626 MB

Scala/Java language

Binary I/O still much faster than text

File Performance - Compression

Binary File	Read Time	Write Time	File Size
Gzip level 6 (Java default)	4 secs	75 secs	286 MB
Gzip level 3	4 secs	20 secs	313 MB
Gzip level 1	4 secs	14 secs	328 MB
LZ4 fast	2 secs	4 secs	423 MB
Raw binary file	1-6 secs	1-6 secs	

Text File	Read Time	Write Time	
Gzip level 6 (default)	26 secs	98 secs	243 MB
Gzip level 3	25 secs	46 secs	259 MB
Gzip level 1	25 secs	33 secs	281 MB
LZ4 fast	22 secs	24 secs	423 MB
Raw text file	18 secs	21 secs	626 MB

Scala/Java language

Binary I/O still much faster than text

File Performance - Compression

Binary File	Read Time	Write Time	File Size
Gzip level 6 (Java default)	4 secs	75 secs	286 MB
Gzip level 3	4 secs	20 secs	313 MB
Gzip level 1	4 secs	14 secs	328 MB
LZ4 fast	2 secs	4 secs	
Raw binary file	1-6 secs	1-6 secs	

Text File	Read Time	Write Time	File Size
Gzip level 6 (default)	26 secs	98 secs	243 MB
Gzip level 3	25 secs	46 secs	259 MB
Gzip level 1	25 secs	33 secs	281 MB
LZ4 fast	22 secs	24 secs	423 MB
Raw text file	18 secs	21 secs	626 MB

Scala/Java language

LZ4 compression ≈ raw I/O speed

File Performance - Compression

Binary File	Read Time	Write Time	File Size
Gzip level 6 (Java default)	4 secs	75 secs	286 MB
Gzip level 3	4 secs	20 secs	313 MB
Gzip level 1	4 secs	14 secs	328 MB
LZ4 fast	2 secs	4 secs	
Raw binary file	1-6 secs	1-6 secs	

Text File	Read Time	Write Time	File Size
Gzip level 6 (default)	26 secs	98 secs	243 MB
Gzip level 3	25 secs	46 secs	259 MB
Gzip level 1	25 secs	33 secs	281 MB
LZ4 fast	22 secs	24 secs	423 MB
Raw text file	18 secs	21 secs	626 MB

Scala/Java language

LZ4 compression ≈ raw I/O speed

File Performance - Summary

- Uncompressed read and write times are comparable
- Binary I/O is much faster than text I/O
- Compressed reads much faster than compressed writes
  - » LZ4 is better than gzip
  - » LZ4 compression times approach raw I/O times