email: abran012@uottawa.ca

SYSC 4700 - Lecture 10
Carleton University

Part 1: *Introduction to Big Data Analytics*

Part 2: *Introduction to Machine Learning*

Presenter: André Brandão
abran012@uottawa.ca
February 08, 2018

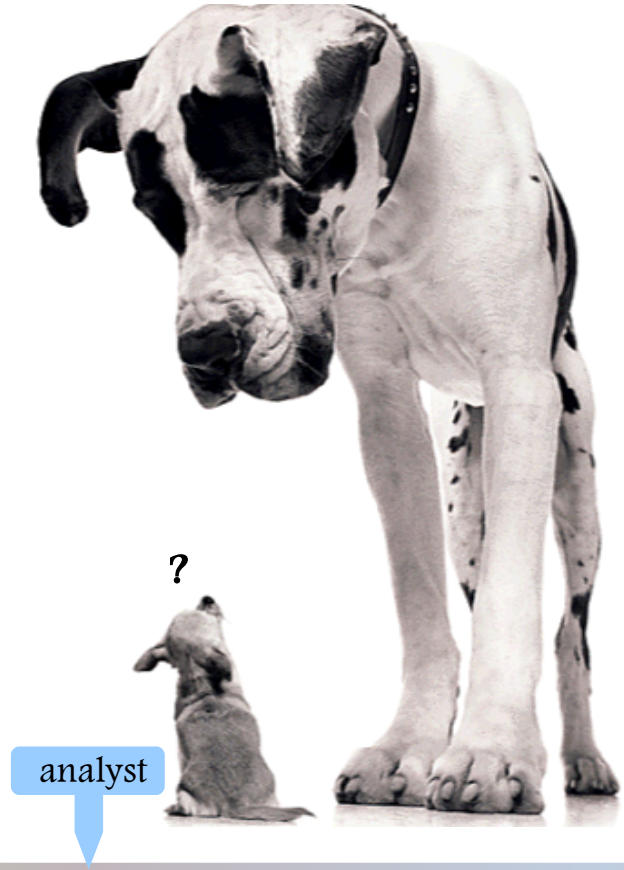
email: abran012@uottawa.ca

SYSC 4700 - Lecture 10 - Part 1

Introduction to Big Data Analytics

We will discuss:

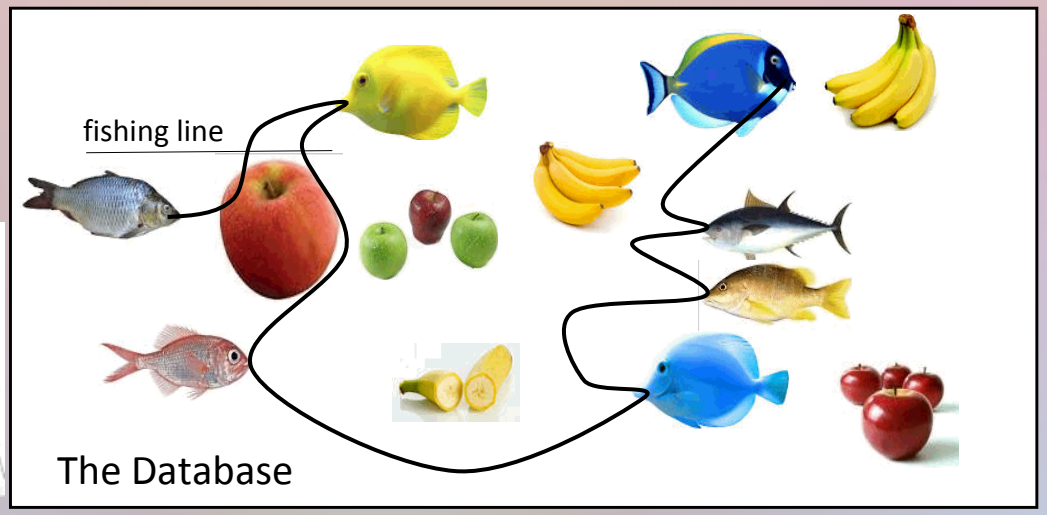
- Basic concepts
- Examples of analytics
- Tapping into the database
- Techniques for data aggregation



email: abran012@uottawa.ca

Basic Concepts

- Data Analytics refers to the extraction of knowledge from large data sets. A large data set contains information about many things.
- The analyst job is to acquire knowledge by linking information that results in a theoretical or practical understanding about something.



email: abran012@uottawa.ca

Knowledge, more than Information

An exercise in analytics: ***W***hat makes a fisherman successful?

You collected lots of data during a fishing season: boat type, nets, bait type, water temperature, wind, cloud cover, water turbidity, time of the day, number of people per boat, jacket colours, fishing rod length, etc. Here you are with a nice table full of information.

Fishing Rod	Temperature	Turbidity	Day time	Bait	Cloud	Wind	Jacket Colour	Fish Caught	Peo
fiber	12	Moderate	07:30:00	fly	30%	5 knots	blue	12	
	10	Moderate	08:15:00	meet	27%	6.6	blue	8	
	10.4	Low	06:30:00	frog	12%	7	red	3	
		High	05:10:00	corn	76%	2	yellow	0	
		High	06:05:00	fly	80%	N/A	blue	7	
			08:20:00	fly	60%	8	red	3	
			17:00:00	worm	40%	12	red	2	
				worm	55%	9	yellow	5	
						9	red		

Table numbers inform that blue jacket fishermen caught more fish than others.

Is that possible fish can see people on the boat when they wear yellow?

email: abran012@uottawa.ca

Your table gets bigger

An exercise in analytics: ***W***ell, you decided to move on and collect data about what is happening on the fishermen's village: the village size, number of churches, stores, theatres, schools, cars, fish processing facilities, transportation, restaurants, sport's arena, community centre, gender distribution, income distribution, family sizes, language and culture in general, etc.

Wind	Jacket Colour	# of Fish	People/boat	Villages	Stores	Churches
5 knots	blue	12	3	6400	8	3
N/A	blue	8	2	9000	12	3
7	red	3	4	568	2	1
2	yellow	6	6	N/A	-	-
8	brown	7	2	125	1	-
...	...	...	...			

You research the stores in small villages selling fishermen's clothing. Affordable clothing are in blue, all the same brand, made in China.

email: abran012@uottawa.ca

Bigger table, bigger picture

An exercise in analytics: ***F***inally, the extra information (i.e. facts) tells you: villagers in that region live within medium to low income communities and buy at local stores. Those who caught more fish wear the same jacket brand from local stores.

Wind	Jacket Colour	Fish Caught	People per boat	Village size	Stores	Churches	Schools	Income Distribution
5 knots	blue	12	6	12000	32	4	5	
06/06/17	blue	8	8	9000	26	6	?	
7	red	3	2	250000	>100	25		
2	yellow	0	1	300000	>100			
N/A	blue	7	5	N/A	N/A			
8	red	3	3	56000	>			
?	red	2	3	98000				
	yellow	5	1	120000				
	red		3					

A second evaluation shows those blue jacket fishermen are locals, from small villages.

You have not concluded what makes a good fisherman yet. You started by noting some blue-jacket guys wearing the same brand, from the same store, catching more fish.

email: abran012@uottawa.ca

Bigger Picture poses a Dilemma

An exercise in analytics: ***A***fter evaluating the local environment as well as all the technical aspects of fishing you encounter a phenomenal Aristotelian dilemma to your analysis:

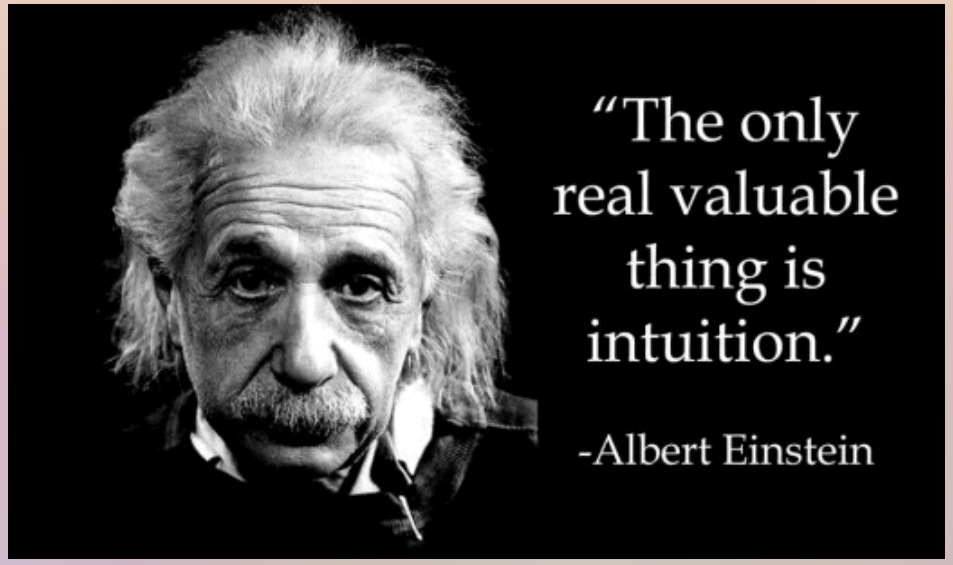
For any hypothesis there can be an infinite number of thesis and no scientific method exists that can find deterministically a thesis that proves a certain hypothesis.

This means: as your database increases, you gain more “insight”, but at the same time the number of variables increase taking with it the number of possibilities for a certain cause-effect event.

email: abran012@uottawa.ca

Analytics and Intuition

An exercise in analytics: ***A***t the end, you use your *Intuition*.



And finally you may conclude:

Local fishermen know best how to catch fish in their own backyards.

email: abran012@uottawa.ca

Modelling and Prediction

We have shown the fishermen story and how the analyst acquired some ***awareness*** about the village and that local environment.

Apart of awareness, the analyst may also ***predict*** behaviour if the amount or "quality" of information in the database is relevant.

The history of ***modelling***, ***prediction*** and ***pattern recognition*** within *Information Theory* is not recent. It gained force with giants like Widrow, Kalman, Abramson and others during the 1960's and matured in the 1980's with the advent of affordable digital signal processors and computers.

Today, the *prediction of customer behaviour is big business in the Internet*. Banks also use prediction techniques for investors in the stock market.

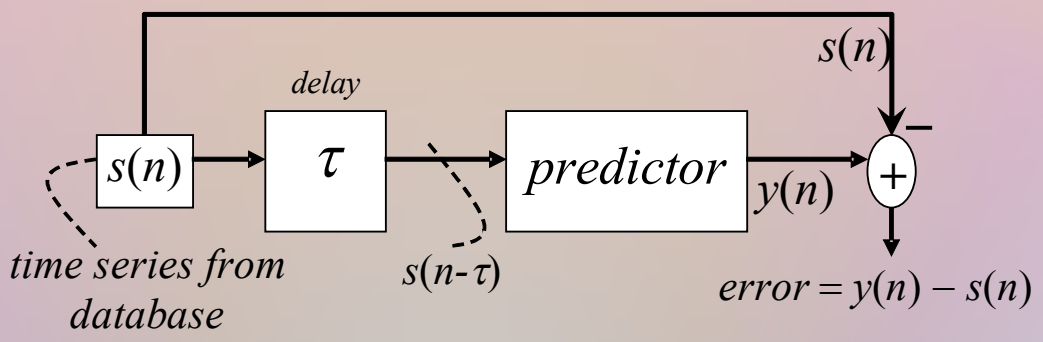
email: abran012@uottawa.ca

Modelling and Prediction

Prediction requires a model. A model in telecommunications is the mathematical transfer function of a system that outputs the desired response when excited with a proper signal.

time series s(n)	
n	*value*
1	0.97
2	0.07
3	0.03
4	0.05
5	0.82
6	0.85
7	0.68
8	0.21
9	0.63
10	0.52
etc...	etc...

A system's engineering view:



When $error \rightarrow 0$ the model reconstructs $s(n)$ by using the information contained in the past series $s(n-\tau)$.

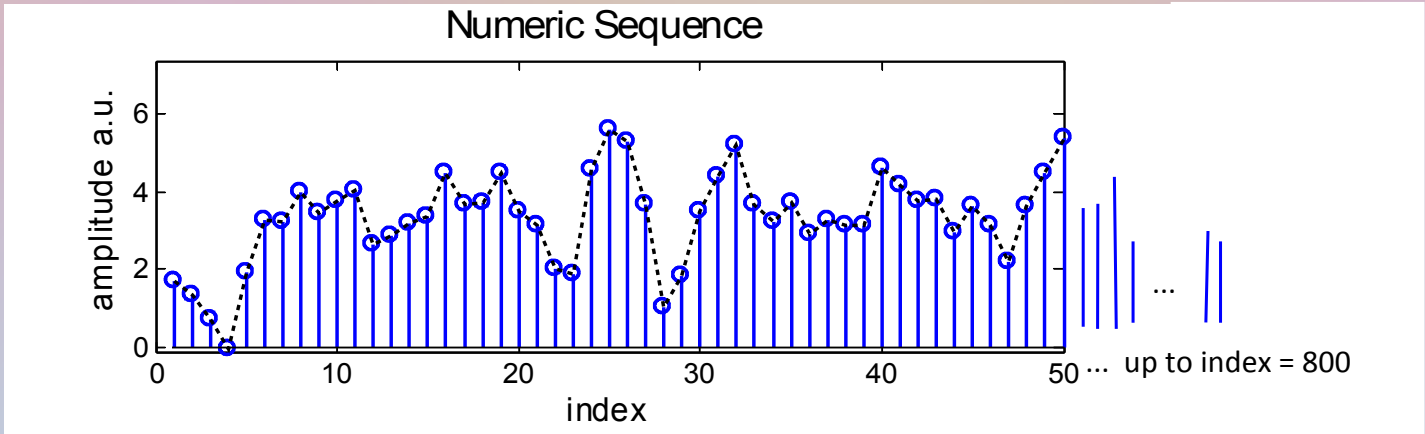
email: abran012@uottawa.ca

Constructing a Model

There are many ways to construct a prediction model for big data analytics. One particular technique, based on the LMS (least-mean-square), is known as the "steepest descent" adaptive algorithm.

The steepest descent is a learning algorithm that adapts its model dynamically as it is fed by a numeric sequence.

For example, consider the following numeric sequence:

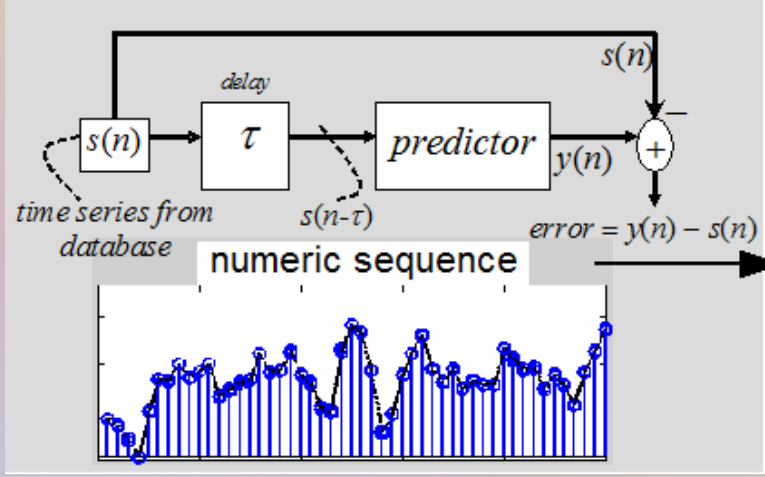


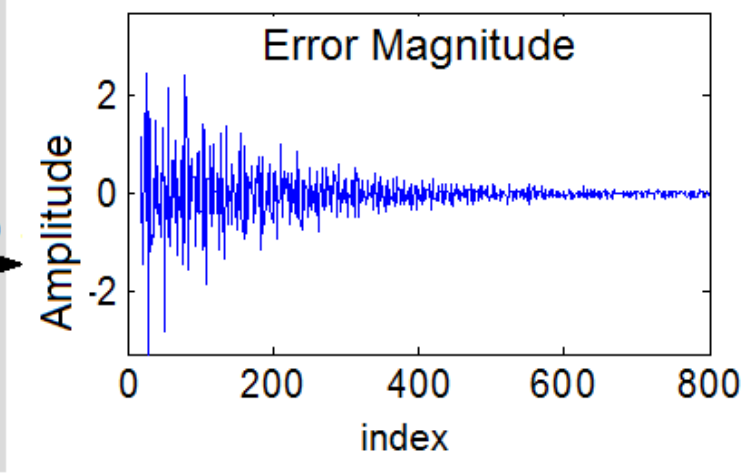
email: abran012@uottawa.ca

Constructing a Model

Given 800 samples (source is your database), what would be the value of sample 801? or 802 and so on.

As the predictor acquires data its engine works out to minimize the prediction error. When the error is minimized the resulting model is ready.





email: abran012@uottawa.ca

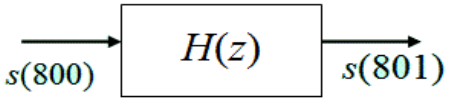
Constructing a Model

The model gives you a mathematical description for the system that generates the original sequence $s(n)$.

In this example our learning algorithm converges into a polynomial which represents the prediction model (transfer function $H(z)$) below:

Coefficient Number	Coefficient Value
1	0.18
2	0.03
3	0.01
4	–0.02
5	0.23
6	0.15
7	0.03
8	0.13
9	0.01
10	0.08
11	0.05
12	–0.13
13	0.02
14	0.02
...	0.02

$$H(z) = \sum_{i=0}^{31} a_i z^{-i}$$



email: abran012@uottawa.ca

Adaptive Linear Prediction

Below is the code that I used for the learning algorithm example. It generates a numeric sequence (suppose this is from a database) and applies the algorithm to find the prediction model $H(z)$. You can play with the parameters and see how the model changes.

```
%Steepest Descent Algorithm
%SYSC4700 Carleton Univ. Lecture on Big Data
Analytics
%A. Brandao - Jan 2017
clear; clf;hold off;
%Generation of a numeric sequence z
% (originally this would come from the database)
N=800; %this is the sequence length
                %Sequence Parameters
a=0.707;  y=0;
x=0.5+0.2*(rand(1,N))+0.7*cos(2*pi*1.3/8*[1:N])+...
0.6*cos(2*3.2*pi/4*[1:N])+...
0.7*cos(2*7.5*pi/4*[1:N])+0.6*cos(2*9.5*pi/4*[1:N]);
```

```
%----- sequence z starts here
for n=1:N,
        z(n)=x(n)+a*y;
        y=z(n);
        intx(n)=n;
end;
%-----end of sequence generation
```

email: abran012@uottawa.ca

Adaptive Linear Prediction

```
figure(1);
subplot(211);
stem(intx,z);
title('Numeric Sequence');
ylabel('amplitude a.u.');
xlabel('index');
%---------------------- Modelling starts here
R=32;        % Steepest descent order
W=[1:R]*0; % Initialize coefficients
X=[1:R]*0;
mu=0.001;
```

```
%Put the engine to work
for n=2:N,
        d=z(n);
        X=[z(n-1),X(1:R-1)];
        E(n) = d - X*W';
        W=W+2*mu*E(n)*X;
        intx(n)=n;
end;
%Vector W contains the prediction model
subplot(212);
plot(intx,E,'b-');
title('Error Magnitude');
ylabel('amplitude a.u.');
xlabel('index');
return;
```

email: abran012@uottawa.ca

Database in the Cloud

Access to big data is normally done in the internet cloud. Many commercial companies offer data repositories over the cloud with servers that respond to SQL (Structured Query Language) enquires from client terminals.

For example, suppose you have data stored in the cloud (i.e. Amazon, Google, etc.) and want to retrieve information about fishermen wearing blue jackets with records of catching lots of fish. SQL is powerful and intuitive. The client would send a message to the SQL database like this:

"Select from fisherman_table where jacket_colour = 'blue' and fish_caught >6"

SQL queries allow you to access big tables and retrieve from the cloud only what is necessary. This saves you local memory space and searching time. An excellent site that allows you to practice with SQL statements is this:

http://www.w3schools.com/sql/sql_select.asp

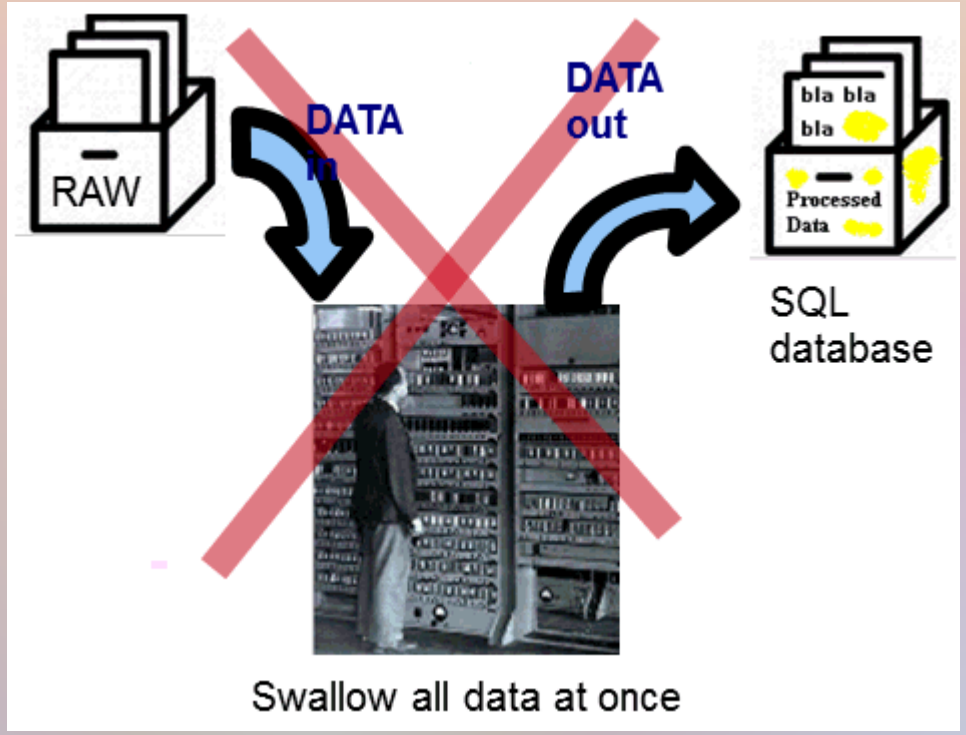
email: abran012@uottawa.ca

Working with Big Data

Although SQL selects only those necessary fields you want from the database, when dealing with large data sets even a few fields may contain more information than the memory of your local computer can handle.

Minimize reading.
(bring data)

Minimize writing.
(push data)



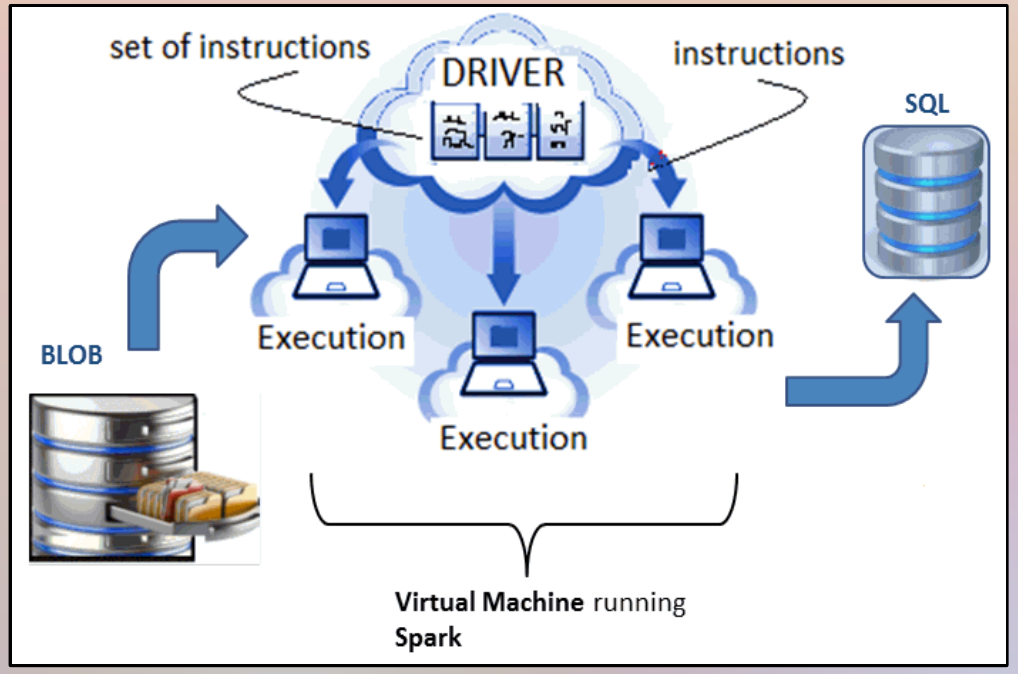
email: abran012@uottawa.ca

The Power of Parallelism

Working with virtual machines (VM), physically located close to your data repository, is the way to go for cluster or parallel computing in the cloud today. Spark is one tool that offers parallelism for big data analytics. Spark is an open source software framework that evolved from Apache Hadoop.

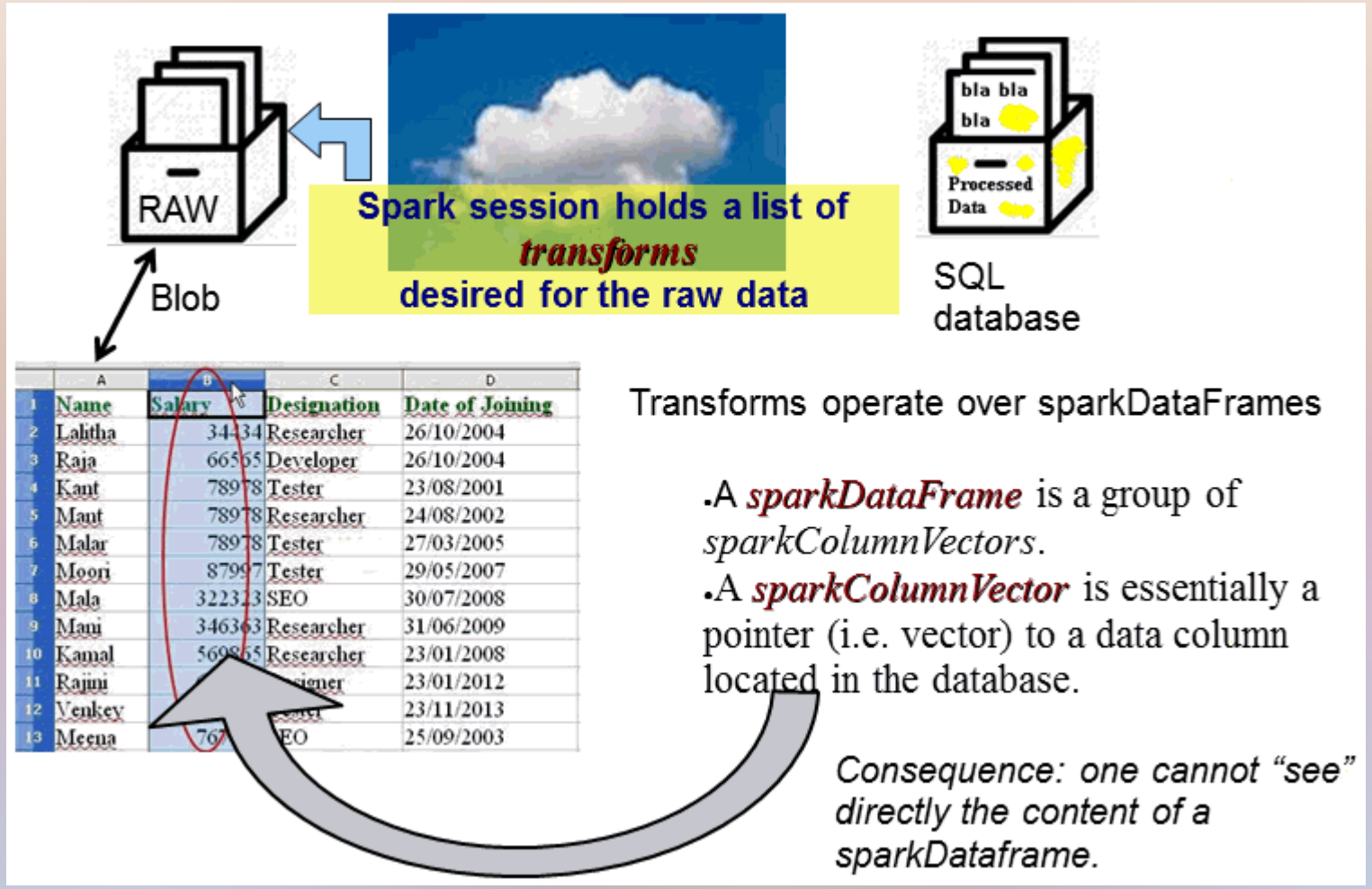
see:

http://spark.apache.org/



email: abran012@uottawa.ca

How Spark Works



	A	B	C	D
1	Name	Salary	Designation	Date of Joining
2	Lalitha	34334	Researcher	26/10/2004
3	Raja	66565	Developer	26/10/2004
4	Kant	78978	Tester	23/08/2001
5	Mant	78978	Researcher	24/08/2002
6	Malar	78978	Tester	27/03/2005
7	Moori	87997	Tester	29/05/2007
8	Mala	322323	SEO	30/07/2008
9	Mani	346363	Researcher	31/06/2009
10	Kamal	569865	Researcher	23/01/2008
11	Rajini			23/01/2012
12	Venkey			23/11/2013
13	Meena	767	EO	25/09/2003

email: abran012@uottawa.ca

How Spark Works

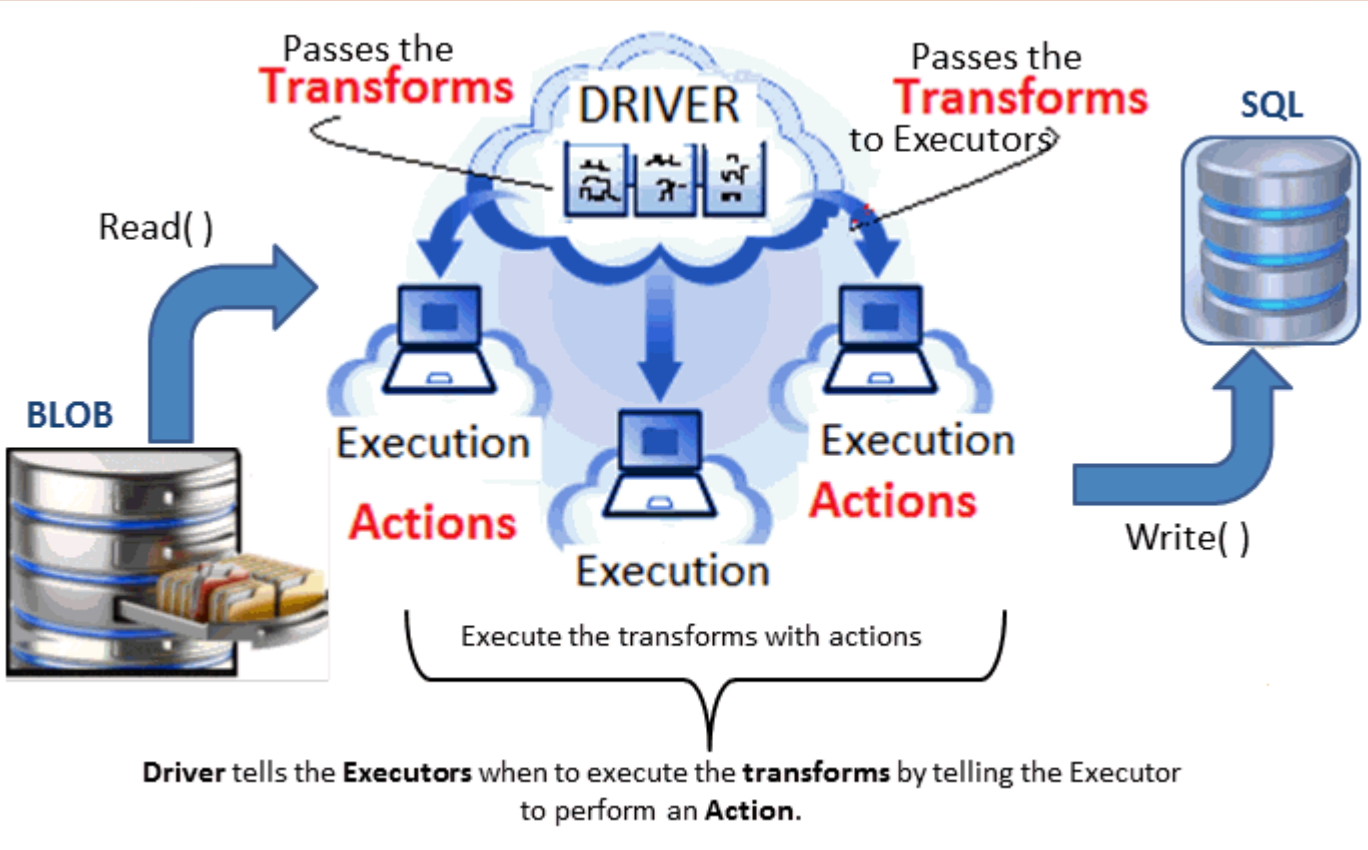
	A	B
1	Name	Salary
2	Lalitha	67,560.00
3	Raja	66565
4	Kant	78978
5	Mant	78978
6	Malar	78978
7	Moori	87997
8	Mala	322323
9	Mani	346363
10	Kamal	569855
11	Rajini	636346
12	Venkey	665656
13	Meena	767403

- The visualization of the content of a *sparkDataFrame* happens after an *action*.
- An *action* is the process of executing a *transform*.

Example: write( ) is an action that results in the physical storage of bits into the database.

email: abran012@uottawa.ca

How Spark Works



email: abran012@uottawa.ca

Concluding Remarks

Big data analytics is part of the digital economy. It is present in all aspects of everyday life. Examples of how diverse this is include:

- Government agencies (statistics, revenue, service delivery, etc.)
- Retailers (understanding customer behaviour)
- Business processes (optimize warehouse stock, deliveries)
- Health care (clustering diseases, etc.)
- Sports (game stats, crowd attendance, etc.)
- Science and research (molecular data analysis, genetics, etc.)
- Safety and Security (clustering for safety and security, etc.)
- Banks (stock markets, etc.)

Apart of Spark there are other frameworks and development platforms devoted to Big Data Analytics (Cloudera, Hortonworks, Databricks, direct SQL, etc.).