

Toronto Apache Spark

Spark & Scala vs The Rest

Tri Nguyen



tritanix@gmail.com
<http://ca.linkedin.com/in/princecharmant>

30 Sept 2015

AGENDA

- Review of the challenge: Air Pollution in Canada
- Review of various implementations:
Relational, Java MR, Pig, Hive, R, Spark
- Live Demo of the Spark Solution
 - Standalone Scala app
 - Interactive scala script
- Why Spark? Why Scala?
- Code & Presentation: <https://github.com/NeuroNex/UG/tree/master/HadoopLab>

ABOUT ME

- Big Data Engineer
- Data Science Certified
(completed Coursera Data Science Specialization with Distinction)
- Hadoop Lead Dev (contract) @ Major Canadian Bank
- Email: tritanix@gmail.com
- LinkedIn: <http://ca.linkedin.com/in/princecharmant>

THE CHALLENGE

QUESTION

- Rank Canadian cities by air pollution level

HOW?

- Get public dataset from National Air Pollution Surveillance (NAPS)
- Use Ozone measures to judge general air pollution level
(more Ozone = more Pollution)

THE DATASET

NAPS Data Products: <http://maps-cartes.ec.gc.ca/rnspa-naps/data.aspx?lang=en>

NAPS Stations (100K, 709 records)

StationID, **STATION_NAME**, Type, Status, TOXIC, Designated, **PROVINCE**, ADDRESS, **CITY**, COUNTRY, FSA, etc...
20101, 56 FITZROY ST., C, 0, , P, PRINCE EDWARD ISLAND, 56 FITZROY ST., CHARLOTTETOWN, CANADA, etc...
30116, HALIFAX CITY HALL, C, 0, , P, NOVA SCOTIA, BARRINGTON & DUKE, HALIFAX, CANADA, etc...
50102, JARDIN BOTANIQUE, R, 0, , P, QUEBEC, BOUL. ROSEMONT, MONTREAL, CANADA, H1X, H1X, etc...
60419, CN TOWER, C, 0, , N, ONTARIO, CN TOWER, TORONTO, CANADA, M5H, , -5, 43.65, -79.38333, etc...

Ozone measurements (10MB, 74064 records for 2012)

PC Stat	YYYYMMDD	AVG	MIN	MAX	H01	H02	H03	H04	H05	H06	H07	H08	...	H20	H21	H22	H23	H24
00701010220120101	27	17	36	17	21	22	29	28	30	31	32	...	29	33	33	34	33	
00701010220120102	30	25	37	33	33	32	31	27	29	29	29	...	33	34	35	37	36	
00706440120120207	-999	28	32	32	31	29	28	28	28	31	30	...	-999	-999	-999	-999	-999	
00706410120120925	9	3	19	11	11	10	7	7	6	3	4	...	10	11	10	9	8	

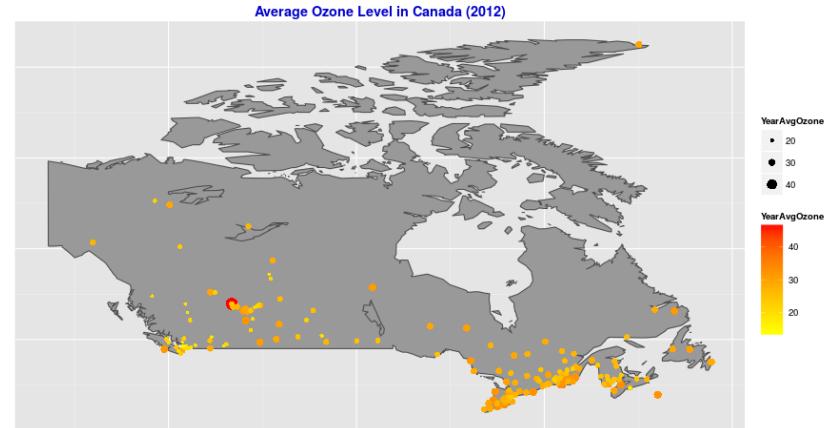
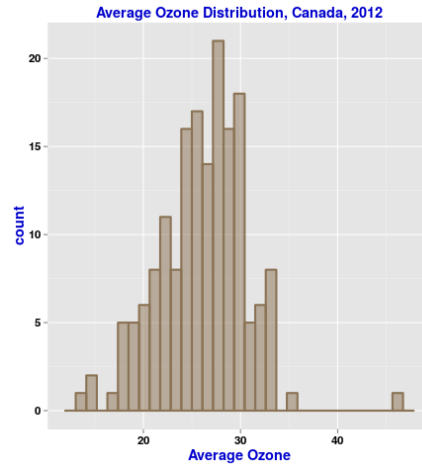
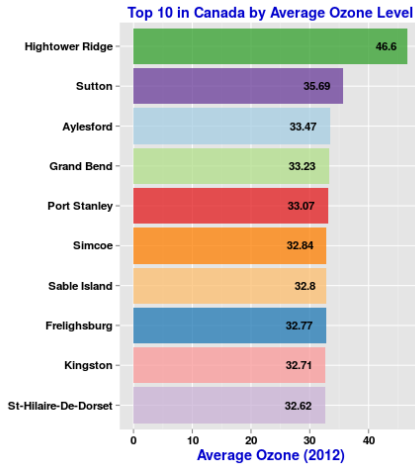
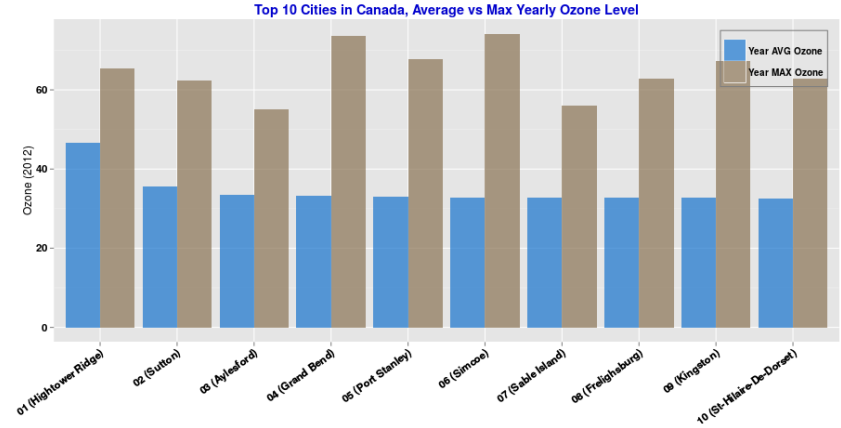
SOLUTION DESIGN

- Per StationID, Per day: **Calculate** Average, Min, Max (**row-wise aggregate**)
- **JOIN**: Ozone Measures, Station on StationID
- **GROUP BY** Province, City: **Calculate** AverageOzone, MaxOzone (entire year)
- **ORDER BY** AverageOzone DESC, MaxOzone DESC
- Save results to CSV

PC Stat	YYYYMMDD	AVG	MIN	MAX	H01	H02	H03	H04	H05	H06	H07	H08	...	H20	H21	H22	H23	H24
007010102	20120101	27	17	36	17	21	22	29	28	30	31	32	...	29	33	33	34	33
007010102	20120102	30	25	37	33	33	32	31	27	29	29	29	...	33	34	35	37	36
007064401	20120207	-999	28	32	32	31	29	28	28	28	31	30	...	-999	-999	-999	-999	-999
007064101	20120925	9	3	19	11	11	10	7	7	6	3	4	...	10	11	10	9	8

EXAMPLE OF RESULTS (by R)

	Province	CityName	YearAvgOzone	YearMaxOzone
1	ALBERTA	Hightower Ridge	46.60118	65.43478
2	QUEBEC	Sutton	35.68524	62.45833
3	NOVA SCOTIA	Aylesford	33.47420	55.16667
4	ONTARIO	Grand Bend	33.23071	73.75000
5	ONTARIO	Port Stanley	33.06979	67.79167
6	ONTARIO	Simcoe	32.83943	74.20833
7	NOVA SCOTIA	Sable Island	32.80191	56.12500
8	QUEBEC	Frelighsburg	32.76678	62.87500
9	ONTARIO	Kingston	32.70555	67.33333
10	QUEBEC	St-Hilaire-De-Dorset	32.61678	62.91667

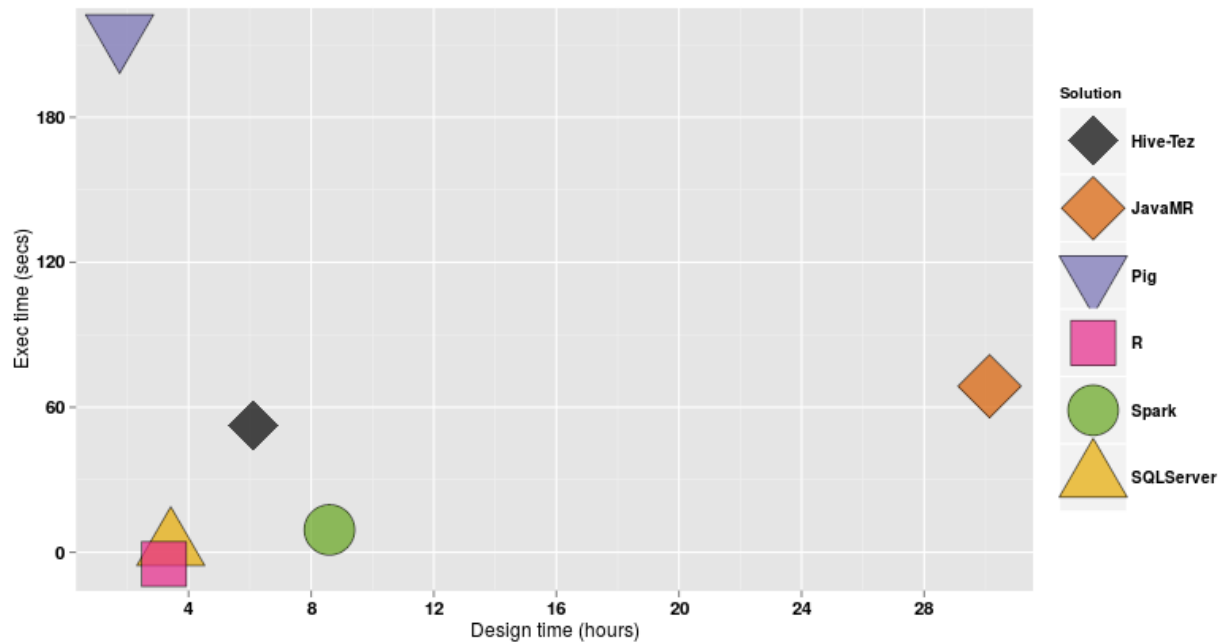


AGENDA

- Review of the challenge: Air Pollution in Canada
- **Review of various implementations:**
Relational, Java MR, Pig, Hive, R, Spark
- Live Demo of the Spark Solution
 - Standalone Scala app
 - Interactive scala script
- Why Spark? Why Scala?
- Code & Presentation: <https://github.com/NeuroNex/UG/tree/master/HadoopLab>

DESIGN COMPARISON

	SQLServer	Pig	Hive (TEZ)	Java MapRed	R	Spark
Exec Time	6 s	3 min 21 secs	60 s	70 s	2 s	9 s
Design Time	4 h	2 h	6 h	30 h	4 h	8 h



AGENDA

- Review of the challenge: Air Pollution in Canada
- Review of various implementations:
Relational, Java MR, Pig, Hive, R, Spark
- **Live Demo of the Spark Solution**
 - Standalone Scala app
 - Interactive scala script
- Why Spark? Why Scala?
- Code & Presentation: <https://github.com/NeuroNex/UG/tree/master/HadoopLab>

WHY SPARK?

	Relational DB	Pig	Hive	Java MR	R	Spark
Programming API (fulfill custom Business Logic)	✓	✗	✗	✓	✓	✓
Big Data	✗	✓	✓	✓	✗	✓
DEV (Design, Maintenance)		💣		💣		
Computation Models: Iterative , Graph	✗	✗	✗	✗	✓	✓
Feature Extension	?	✗	✗	? (★)	∞	SparkSQL Streaming ML

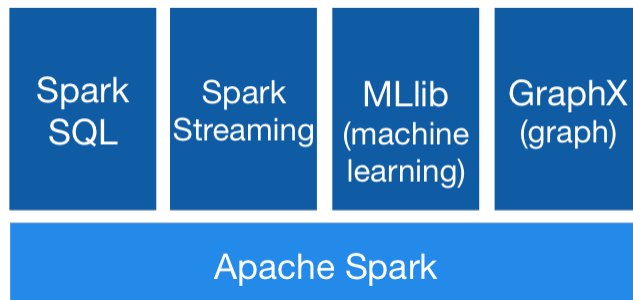
(★) “spark is already replacing mapreduce for most new applications. but mapreduce won't die”

Doug Cutting, Cloudera Webinar [Uniting Spark and Hadoop: The One Platform Initiative](#) (2015-09-24)

WHAT IS SPARK?

- Fast and General engine for large-scale data processing
 - Combine SQL, streaming, and complex analytics
 - Multiples computing models: MR, Iterative, Graph
- Java, Scala, Python, R
- Access data on: HDFS, S3, Tachyon, HBase, Cassandra, JDBC, etc.
- Run on: Hadoop, Mesos, Standalone, Cloud

[Apache Spark Web Site](http://spark.apache.org)



WHAT IS SCALA?

- [Quick Intro \(scala-lang.org\)](http://scala-lang.org)
- [More developed intro \(Wikipedia\)](#)
- Scalable (**SC**ale **L**anguage)
- Address shortcomings of Java while still 100% compatible Java bytecode
- Full interoperability with Java
- Expressive syntax (more concise, less verbose)
- Object Oriented
- Functional
- Adoption: Spark, Twitter (Kafka, Samza), Apache Ignite

IS SCALA SIMPLE OR COMPLEX?

- Minimal Syntax != Simple
 - lots of simplifications done behind the scene to simplify developer task:
implicit class, type inference
 - Need to be aware of too much automation Scala compiler cannot always guess what you want.
- Support both OO & Functional paradigms == ++Complexity
- High Learning curve but worth the investment
- IDE support for Scala is not (yet) as convenient as for other languages

WHY SCALA?

- Scala vs Java: [How-to: Run a Simple Apache Spark App in CDH 5](#)
 - Scala: 25 lines
 - Java : 75 lines
- [Why Should I Learn Scala?](#)
- [What do you think about the Scala programming language?](#)

LEARNING

- Book: [Learning Spark](#)
- [Scala Documentation](#) (scala-lang.org): Docs, FAQ, Tutorials, Tour
- Apache: [Scala API](#), [Spark API](#), [Spark Programming Guides](#)
- Communities: [StackOverflow](#), [Databricks forum](#)

ONLINE COURSES

- Udemy: [Introduction to **Apache Spark** for Developers and Engineers](#) (Scala)
- edX: [Introduction to Big Data with **Apache Spark**](#) (Python)
- UC Berkeley: [AMPCamp Big Data Bootcamp](#)