Philippe ROSSIGNOL

Big Data Consultant: Architect, Data Engineer, Certified in Data Science

Communicating with an Agile Culture

Approache driven Use Cases and Data

Prepare the Data, create the Dashboards

Certifications in Data Science

Domains: Bank, Insurance

LinkedIn Github

WordPress

Driving License

Bordeaux (33000) France

prossblad@gmail.com

+33 6 08 78 77 85

pjlrossg

Experiences

Big Data Manager

AKKA Technologies - Since January 2017 - Niort - France

In charge of the Big Data offer for AKKA Technologies group

Spark on Hadoop Expertise - Data Engineering, Architecture and Teaching

Spherea - December 2016 to January 2017 - Freelancer - Toulouse - France

- Expert around Spark architecture on top of Hadoop. Taught Spark developments (with DataSets and RDDs) by using MongoDB, CSV, JSON and XML libraries. Taught agile prototyping with Notebooks such as Jupyter and Zeppelin.
- Installation and configuration of a plug-and-play Datalab environment enabling executions on multiple different clusters whatever the target Hadoop distribution types (Hortonworks, Cloudera, MapR or pure Hadoop). Integration of the latest Spark V2.1.0 and Spark history server, testing on Hortonworks V2.5 (in AWS and Local modes). Preparation of open source tools for Data-Engineers, Data-Analysts and Data-Scientists (e.g: Jupyter Notebook and Zeppelin for running on top of Spark). Docker containerization for Hadoop nodes, Spark applications and Notebook servers.

Big Data Expertise - Data Engineering and Architecture

COVEA - January 2016 to July 2016 - Freelancer - Niort - France

- Data Engineering and Big Data Architecture projects driven by Data and Use Cases:
  Data transformations with Spark (e.g: XML and log4j logs to column format), Dashboards with Banana and SolrCloud search engine, Machine Learning with Spark ML, Prototyping with Jupyter Notebook and Spark kernel, On-demand deployments for Spark applications, Developments with Python Java and Scala, Trainings.
- Quick delivering of high challenge projects oriented logs analysis:
  - Leading a Data driven approach around the EDRMS (Electronic Document and Records Management System), by implying closely the project's owner in an agile way in order to clarify his needs like getting statistics and detecting outliers.
  - One of the challenge was to transform a big variety of XML logs coming from the EDRMS via a Java Spark transformer, to make complex aggregations with Spark SQL, then to perform indexing with Solr-Cloud, and to finally for analyzing of aggregated clean data with Banana dashboards.
- Machine Learning POC for quality analysis of CRM data:
  Using Spark ML and GraphLab-Create for researching hidden abnormalities in the data. Data preparation, and clustering algorithms such as K-Means, GMM and LOF.

Big Data Expertise: Data Engineering and Architecture

BNP Paribas Cardif Insurance - November 2014 to December 2015 - Freelancer - Nanterre - France

- Development, Analysis, Administration, Architecture:
  Innovating and assisting the Datascientists who develop their business Use Cases with Machine Learning by using the benefits provided by Big Data technologies such as Hadoop with Spark, with DataFrames and Machine Learning algorithms (e.g of UCs: Anti-Fraud detection, churn, appetence, text analysis for classification).
- Recommendation system:
  - Implementation of a Proof Of Concept consisting of a real time recommendation system to recommend insurance products on new clients's sales receipt when they go through the tills of hypermarkets.
  - Spark Machine Learning Pipeline for learning in batch mode: Supervised learning with the Alternative Least Square of Spark-ML (matrix factorization algorithm, using the vector of users and the rank parameter for dimensions reduction).Unsupervised learning with the K-Means algorithm of Spark-ML in order to find clusters of similarities in term of purchase behavior.

- Hadoop HDFS for storing logs coming from the tills in order to process later feature engineering and learning algorithms. Hadoop YARN for Spark and Kafka clusters (for processing and computing in batch mode and in real time mode).
- Real time with Kafka and Spark Streaming for predictions (evaluations of sales receipts coming from the tills of supermarkets).
- Spark Python programming. Developing a Python-Scala bridge in order to improve the performances of UDF (User Defined Functions). Linux shell programming for packaging and deployment in production environment. Jupyter Notebook and Eclipse-PyDev IDE used as development environments.

- PageRank:
  - Understanding of the PageRank algorithm and developing it with Spark RDDs in order to make a pedagogical presentation of Graph Processing for the Datascientists and to convince them to use Spark GraphX coupled with a Graph database like Neo4J.
  - Implementation of Proof Of Concept based on Spark RDDs (to present the PageRank algorithm) and Spark GraphX (for PageRank, Connected Components, and Triangle Counting). Tehnos: Spark / Hadoop YARN, Hadoop HDFS, Python, Scala, Jupyter Notebook.
- POC with Spark DataFrames (SQL) and Spark Machine Learning:
  - Implementation of Proof Of Concepts with Spark DataFrames and SQL, and Spark Machine Learning (by using objects like Transformer and Estimator for Pipelines, Evaluator, CrossValidator).
  - Developments by using ML algorithms like the Linear and Logistic Regressions, Random Forest, Neural Networks and ALS for recommendations.
  - Presentation of these works to the DataScientists in order to explain them how from their Jupyter environment they can develop both with Python Scikit-Learn and Pandas and with Spark ML and DataFrames.
  - Technos: Spark ML, Spark DF and SQL, Jupyter Notebook, Python, Pandas.
- Bench project – Pig Hive vs Spark:
  - Implementation of a bench in order to compare performances between Pig Hive and Spark SQL on a five nodes Hadoop V2.6 cluster, based on a Left- Outer Join query with tables containing retail data.
  - Technos: Shell Linux programming for packaging and deployment of the project on different Hadoop-Spark clusters, Spark DataFrames, Python.
- Development of Csv2Hive:
  - Development of an injector named Csv2Hive available on GitHub https://github.com/enahwe/Csv2Hive.
  - This tool infers dynamicaly the schema from big CSV files containing lot of columns; this tool enables quick automatic injections of external data to feed Hive metastore and Hadoop HDFS.Technos: 95% of Linux Shell scripting, 5% of Python.
- Miscelaneous tasks:
  - Administration of a Cloudera 4 nodes cluster (Hadoop 2.3.0-cdh5.0.2) for using mainly Pig and Hive.
  - Installation and administration of a 5 nodes Cloudera cluster (Hadoop 2.6.0- cdh5.4.4), Sizing for each node: 4 cores 2.6 Ghz, 96 GB Mem, 1 TB Disk.
  - Installation of Spark on YARN with Anaconda on each Hadoop node.
  - Configuration of Jupyter Notebook for Spark on YARN, to allow DataScientists to discover Spark in Hadoop cluster.
  - In charge of feeding of business data towards the HadoopDataLake (hence Csv2Hive).
  - Developed MapReduce jobs in Java (e.g: inverted index).
- Machine-Learning Challenge (Retail domain):
  Multi-categorization for CDiscount company (challenge on https://www.datascience.net/fr/challenge/20/details). Developed a program with more than 500 Multinomial-NaiveBayes models, Stemming, Stratified sampling and Mutual Information.

Consultant at BULL - Missions for CDISCOUNT

CDISCOUNT - May 2014 to September 2014 - Consultant - Bordeaux - France

- 09/2014 - Bench for Solr search engine:
  Defining a Test Plan to compare performances between the two search engines Solr and Exalead.
- 08/2014 - Call of tender:

- Survey around a possible hybrid SQL-NoSQL implementation for a multilingual platform based on SQL-Server and Cassandra.

- 05/2014-06/2014 - Bench for Kafka broker:
  - As part of the implementation of a strategic large project, defining a performance Test Plan based on Apache Kafka broker used to feed the NoSQL Cassandra database.
  - Development of a configurable tool (production in real-time of big volumes of data with statistics, by using customized message formats as Text, Json, Xml and others) covering all the test cases for Kafka and Cassandra in the new Cdiscount environments.

Consultant at BULL - Big Data Proofs Of Concepts

BULL - March 2014 to April 2014 - Consultant - Bordeaux - France

Big Data implementations:

- Proofs Of Concepts based on Hadoop Cloudera distribution, with recommenders based on Naïve Bayes (Python, Apache Mahout and Java).
- Proofs Of Concepts based on Zookeeper, Kafka, Storm and Cassandra.
- Survey to combine M2M services into a Big Data infrastructure.
- Machine Learning: Demos around Neural Networks for Classification (supervised learning).

Consultant at BULL - Mission for CETE

CETE - January 2014 to February 2014 - Consultant - Bordeaux - France

Traffic jam web application:
Architecture surveys around a web application implemented in Java and JavaScript, dedicated to produce contents in real time about the traffic jams, car crashes, roadwork's, etc.

Consultant at BULL - Mission for VOYAGES-SNCF

VOYAGES-SNCF - September 2013 to December 2013 - Consultant - Nantes-Lilles - France

Yield Management in the domain of train transportation:

- Audits for designing n-tiers architectures of existing applications dedicated to the Yield Management business domain (optimizing the prices for train tickets).
- List of technologies involved: AngularJS, JQuery with JQPlot for the GUIs, web-services with JBoss AS server configured in High Availability, Drools Engine Rules (BRMS) to compute automatically the recurrent and simple business rules.

Consultant at BULL - Architecture around Banking Platform

BULL - July 2013 to August 2013 - Consultant - Bordeaux - France

Investigations around a supervision platform for banking terminals:
Audits around new scenarios and new technologies in order to evolve the existing platform and to boost the performances and scalability.

Consultant at BULL - Mission for POLE-EMPLOI

POLE-EMPLOI - May 2010 to June 2013 - Consultant - Bordeaux - France

- Support and compliance of operational processes and action plans:
  - Compliance of the project plans and operational processes to successfully deliver in time the statistical business applications
  - List of technologies: Java-EE, Customer's frameworks, ClearCase, Maven, SonarQube, WebLogic, Oracle, SAS, Unix.
- Design and development of a SSO launcher for SAS Enterprise Guide V4.3:
  Launcher deployed in all the agencies, providing an automatic authentication for the end-users like the statisticians.
- Design and development (in Java) of a reliable integration chain for SAS components running in Cobol and Unix:

- Solution similar to a continuous integration system, automatically preparing each SAS component for a specific target environment such as the production.

Consultant at BULL - Pre-sale for CITY HALL OF BRUSSELS

CITY HALL OF BRUSSELS - April 2010 - Consultant - Bordeaux - France

Quotation for a call for tender successfully won, consisting of the creation of a Web Java EE Application for managing the European historical heritages (method used: Use Case Points).

Consultant at BULL - Mission for POLE-EMPLOI

POLE-EMPLOI - January 2010 to March 2010 - Consultant - Bordeaux - France

Audits around the Business and Project owners:
Definition of a specification to start as soon as possible a new web Java EE application in order to facilitate the search of jobs.

Consultant at BULL - Mission for CETE

CETE - January 2009 to December 2009 - Consultant - Bordeaux - France

Designing new architectures:
Specifications around Java ESB technologies in order to transfer the business data in a secure way and reliable way (solution based on Java OSGI and Apache Camel).

Consultant at BULL - Mission for MAAF

MAAF - November 2007 to December 2008 - Consultant - Niort - France

Designing architectures:
Opportunity surveys, audits, feasibility surveys, costings.

Consultant at BULL - R&D European Project Manager

ITEA2 - December 2006 to October 2007 - Consultant - Bordeaux - France

French leadership and coordinator of a R&D European project called Usenet (ITEA2 consortium), in order to create a new European standard in the domain of the M2M (Machine to Machine).

Consultant at BULL - ETL Projects & Pre-sale

BULL - September 2006 to November 2006 - Consultant - Bordeaux - France

- 11/2006 - Architecture and Development around ETL (Talend) for COMPLETEL-BOUYGUES:
  Design and development of a Java ETL application based on Talend, in order to transfer monthly and automatically the invoices which come directly from the clients.
- 10/2006 - Architecture and Development around ETL (Talend) for CNAMTS:
  Design and development of a Java ETL application based on Talend, in order to import-export the data from the fleet management sources, to SIEBEL and GLPI.
- 09/2006 - Pre-sale & Investigations:
  - Costings and technical surveys around solutions using GPS localization in the domain of Fleet Management.
  - Investigations to make recommendations in order to evolve client-server architectures to N-tier JEE architectures.
  - Investigations around access control systems concerning the time management.

Java-EE Expertise - Development and Architecture

LECTRA - September 2002 to August 2006 - Full-time - Bordeaux - France

- Technical support in order to help the developer teams.

- Development around a Java ETL (based on Oracle Sunopsis) for real-time synchronization and bidirectional communication between heterogeneous databases.
- Specifications and developments in Java around a web PDM application (Product Data Management). List of technologies: Rational Rapid Developer, WebSphere, Tomcat, Oracle RDBMS, tests with IBM Workload Simulator.
- Development in order to install automatically any Oracle 10G Database in "silent" mode. List of technologies: InstallShield, Java, Ant.
- Many developments based on Java EJBs, JBoss, WebLogic and WebSphere servers.

Java Expertise - Development

BULL - September 1997 to August 2002 - Full-time - Bordeaux - France

- Back-office Java developments for Call-Centers.
- Development in Java of a CTI server (Computer Telephony Interface) for Alcatel-Lucent PABXs, above Genesys middle-ware, providing a CTI integration with high availability up to hundreds connected operators using a CRM application (Customer Relationship Management).
- Main technologies: Java, Siebel CRM, CTI Genesys, Oracle RDBMS, MySQL RDBMS, SQL-Server RDBMS, MQSeries Broker, SWIFT, LDAP Directory.
- Main customers: CNAMTS, MGEN GROUP CIC, GROUPAMA, URSSAF, CNCA.

Electronic Expertises

MATTHEWS SWEDOT, GAIA, IBM, SOULE - January 1990 to August 1996 - Full-time - Bordeaux - France

- 01/1995-08/1996 - Electronic technician at MATTHEWS SWEDOT:
  In the domain of industrial printer systems, maintenance, after-sales and trainings.
- 06/1994-08/1994 - Electronic Engineer (trainee position) at GAIA:
  Design and creation of miniaturized power supplies based on high frequencies.
- 01/1993-08/1993 - Electronic technician at MATTHEWS SWEDOT:
  Maintenance and after-sales in the domain of the industrial printer systems.
- 1992 - Mainframe systems tester at IBM:
  Technician system tester for Mainframes IBM 3090 and ES9000.
- 1990-1991 - Electronic technician at SOULE:
  Design, implementation and testing lightning for detection systems (customer: Electricite De France), based on Motorola microcontrollers.

Skills

Big Data: Data Engineering

- Data Transformations on big volumes: Cleaning, Feature Engineering, make the data ready for statistical analysis and Machine Learning, in batch and real-time
- Unstructured Data to Structured Data (e.g: XML and Logs transformations to columns formats)
- Spark / Hadoop YARN (Dataframes, RDDs, Streaming) and Hadoop HDFS
- Hadoop MapReduce/Tez with Hive & Pig
- Ingestion: Kafka and RabbitMQ with Spark, Sqoop

Big Data: Data Visualization

- Dashboards Banana (fork of Kibana) / Solr-Cloud
- Dashboards Hue-Search / Solr-Cloud
- Search Engine: Indexing the transformed and cleaned data with Solr-Cloud

Big Data: Data Science

- Spark algorithms: Regression & Classification, Clustering and Recommendation

- Using Spark ML with both Dataframes and RDDs
- Scikit-Learn & Pandas
- Turi (GraphLab Create)

Development (Big Data and Back Office)

- Languages: Java EE, Python
- IDEs: Eclipse, IntelliJ, Jupyter Notebook, Spark Notebook, Apache Zeppelin
- Linux Command Line and Shell

Architecture (Big Data and standard systems)

- Hadoop and its whole Ecosystem
- Cloudera
- Java Application Servers & RDBMS
- Applicative & Physical architectures

Spoken languages

- French ★★★★★
- English ★★★½☆

Education

Machine Learning Regression (6 weeks) - Certification (100%)

University of Washington (MOOC Coursera)
February 2016 to April 2016

Regression is one of the most important and broadly used machine learning and statistics tools out there. It allows you to make predictions from data by learning the relationship between features of your data and some observed continuous-valued response.

This course covers the major topics such as:
Simple Linear Regression.
Multiple Regression.
Assessing Performance.
Ridge Regression.
Feature Selection & Lasso.
Nearest Neighbors & Kernel Regression.

This course teaches how to:
Describe the input and output of a regression model.
Compare and contrast bias and variance when modeling data.
Estimate model parameters using optimization algorithms.
Tune parameters with cross validation.
Analyze the performance of the model.
Describe the notion of sparsity and how LASSO leads to sparse solutions.
Deploy methods to select between models.
Exploit the model to form predictions.
Build regression models to make predictions.
Implement these techniques in Python.

Machine Learning Foundations: A Case Study Approach (6 weeks) - Certification (100%)

University of Washington (MOOC Coursera)

January 2016 to February 2016

Implementing intelligent applications using regression, classification, nearest neighbor search, clustering, collaborative filtering, and deep learning... With these ML methods, intelligent applications can perform predictions, personalized recommendations and retrieval, learn non-linear features that improve the accuracy of the solutions, and much more.

This course covers the major topics such as:
Regression.
Classification.
Clustering & Retrieval.
Recommender Systems & Dimensionality Reduction.
ML Capstone: An Intelligent Application with Deep Learning.

This course teaches how to:
Become a machine learning expert, ready to develop and deploy new intelligent applications.

Machine Learning (11 weeks) - Certification (100%)

Stanford University (MOOC Coursera)

November 2015 to January 2016

Linear Regression with one and multiple variables, Gradient Descent and Cost Function, Linear Algebra, Logistic Regression, Regularization, Multiclass Classification, Solving the Problem of Overfitting, Neural Networks and Backpropagation, Evaluating a Learning Algorithm, Bias vs. Variance, Support Vector Machines, Unsupervised Learning, Dimensionality Reduction, Anomaly Detection, Recommender Systems, Density Estimation, Multivariate Gaussian Distribution, Collaborative Filtering, Low Rank Matrix Factorization, Gradient Descent with Large Datasets, Photo OCR.

Big Data Science Training - Spark for Data Scientists (2 days)

Xebia Training - Paris

September 2015

“Hands-on” training focusing on Spark Machine Learning and DataFrames in order to develop Machine Learning Pipelines (e.g: with objects like Transformers, Estimators, Cross-Validators and others.).

Big Data Training - Spark for Developers (3 days)

Xebia Training - Paris

June 2015

Cloudera-Xebia three days Spark course enables participants to build complete, unified Big Data applications combining batch, streaming, and interactive analytics on all their data. With Spark, developers can write sophisticated parallel applications to execute faster decisions, better decisions, and real-time actions, applied to a wide variety of use cases, architectures, and industries.

Big Data Training - Introduction to Data Science with Hadoop (4 days)

ExitCertified - Montreal

April 2014

Introduction to Data Science and Machine Learning (Clustering, Classification and Recommenders). Creating of Hybrid Recommenders with Apache Hadoop, Apache Mahout, Hive, and R by using algorithms such as Naïve Bayes, Tanimoto coefficient, Euclidean distance, etc. Labs were based on a fictive movies company use case, similar to Netflix.

Big Data Training - Developer for Apache Hadoop (4 days)

Xebia Training - Paris

March 2014

MapReduce development jobs with Hadoop (in Java and Python)

Business Intelligence SAS Trainings - The whole SAS Platform

SAS Institute

January 2011 to September 2011

SAS Administration and SAS Developer trainings:
SAS Programming Levels I and II, Macro SAS Language, SAS Java API, SAS Enterprise Guide GUI, Architecture and components of the SAS platform, Administration of SAS platform

Master Diploma - In Computer Science and Development

ENST - Brest (FRANCE)

September 1996 to September 1997

Micro-Electronic Diploma - NVQ Level 5

I.X.L Laboratory - University of Bordeaux (FRANCE)

September 1993 to 1994

Micro-Electronics

Electronic Diploma - Higher Education

Night School - University of Bordeaux (FRANCE)

September 1992 to 1993

Electrical and Electronics Engineering

Electronic Diploma - BTEC Higher National Diploma

Night School - University of Bordeaux (FRANCE)

September 1990 to 1992

Electrical and Electronics Engineering

Electronic Diploma - NVQ Level 3 Diploma

AFPA - Pau (FRANCE)

September 1988 to 1989

Industrial Electronics Technology/Technician

Interests

Miscellaneous

- Sports: Nature & Surfing (7"6)
- Cooking: Healthy food (spicy)
- Books: Science, Philosophy, Nature medicine
- Movies: Fantastic, Humouristic, Cartoons