Big Data & Data Science Architect

PERSONAL INFORMATION

Name: João P. A. Cerqueira

Sex: Male Date of birth: 19/12/1983

Address UK: 21, Woodlands Avenue, Worcester Park, KT47AL

Contacts: (+44) 07572550311

Email: jpacerqueira@gmail.com

Skipe: jpacerqueira83

Linkedin: uk.linkedin.com/in/joaopedroafonsocerqueira

github: github.com/jpacerqueira

personal site : https://fuelbigdata.com

WORK EXPERIENCE

- Lead Data Consultant - Data Science and Data Engineering - GFT Financial Limited

(From June2019 - current Date)

GFT Financial Limited - City of London (**https://www.gft.com/uk/en/index/**) (Permanent)

Business or sector: Consultancy services via GFT Financial Limited UK, for non-financial services customers.

Technologies used:

1. Design and implement an Multi-Cloud solution for Data Science Analytics, fous in Python solutions.
2. Setup of a Docker Hub solution for Data Science using :
    1. Python 2.7 - 3.6
    2. H2o.ai free tier
    3. Spark latest version up to 2.4.x
    4. Establish of Prototypes as consumable Docker Hub images, valid in Multi-cloud and private Cloud.
    5. Establish a public contributor repo for democratic usage of the solution.
        1. Detials in https://github.com/jpacerqueira/Jupyter_Spark_H2O_Kafka_Client_Setup
3. Establishment of my consultancy in GFT using the principles of Fuel BigData .
    1. An entity dedicated in Data Science consultancy.
    2. Applied Full Stack Data Science with Business Intelligence, Analytics and Machine Learning tools, principles and solution definitions.
    3. Follow site for more details https://fuelbigdata.com .

Project roles :

1. Got involved in new customer engagement, supporting new sales pitch for new markets. Establish principles and define general technical solution approach for Data Science Multi-cloud (GCP, Azure, AWS)
2. Lead Data Architect for the 1st UK customer outside Financial Services, the customer is

"undescribed in retail Property Sector" . Establish the relationship with the customer with their new Technology department. Bring a foundation solution for their requirements around Business Intelligence and Data Collection at the business level. Implement the architecture solution in Azure Cloud . Define A new Solution for data consumption and servicing using Azure Citus PostGreSQL to collect, exchange and service information to partners, via a WebApp solution. Define a General Approach for Analytics for such application, with ETL support using Python and PySpark with Apache Arrow accelerators. Solution at TPOC level, was handled for development with GFT nearshores.

- Big Data & Data Science Architect - Head of Big Data - Perform Media Group

(From November 2015 – June2019)

DAZN Media Group (ex- Perform Media) - London (**http://www.performgroup.com/contact/**) (Permanent)

Business or sector: Perform Group is a Sports Media group owner of Brands: Opta, Runningball, Goal.com, SportingNew.com and DAZN.com (OTT b2c video subscription).

Technologies used:

4. Design and implement the Big Data full stack at Perform Group using the following technologies:
    1. Cloudera Enterprise Data Hub (PaaS) , CDH 5.9 and recently CDH 5.14 .
    2. Oracle enterprise Big Data Appliance (BDA) , with Cloudera CDH.
    3. Implemented the new generation of IOT collection from tracking data supply of tracking cameras to enrich Opta event data. Usage of Akka collection tools using Fast Data principles and Kafka buffer storage for Spark Streaming analytics.
    4. Implemented a "Single Customer View" across the Portals of Perform Group , goal.com sportingnew.com , dazn.com , dabblebet.com and others to exchange customer profiles and unique_id between DW, Marketing Cloud CRM and individual site profiles. Designed a solution using Neo4J for graph calculation and Big Data Cloudera CDH for log analysis.
5. Design of a new AWS Cloud solution for DW for DAZN product using AWS EMR, RedShift, RedShift Spectrum, AWS Glue Crawlers.
    1. Design the new generations of S3 lake based on Object Store.
    2. Design the new generation of MetaStore collection using Schema on read over object store with AWS Glue Crawlers.
    3. Enable the reception of Media Portal profiles and customers in the new S3 Lake, from the on-premise solution designed as Single Customer View for Perform Media Services.
6. Designing a New general framework for ETL Solutions based on:
    1. Control of /raw /staged /published areas
    2. Utilize ETL technologies Apache Spark 1.6.0/2.1.0/2.3.0 (Python and Scala)
    3. Achieve published insights for apache Hive.
    4. Served insights , on-premise using Oracle Big Data SQL.
    5. Served insights, in-cloud AWS with using RedShift Spectrum.
7. Full setup of Data Science Stack for Brands Opta and Perform Media
    1. Used technologies based on Python and R languages for data science.
        1. R-Studio and R as a data science stack of the new generation of Perform logic.
        2. Setup of services with Oracle R Enterprise (ORE dbi) for ORACLE DB.
        3. Setup for BigData with Hadoop/Hive using ORCH R/cran.
            1. Design of processes and setup of services for the Data Science team, to align in the SDLC of development teams.
            2. Setup of data science services with sparklyr and tools of a new generation of machine

learning and Deep learning services using R H20.ai .

4. Setup for BigData with Hadoop/Hive using PySpark.
    1. Design of processes and setup of services for the Data Science team, to align in the SDLC of development teams.
    2. Setup of DataScience stack with pyspark, sparkling water, h2o.ai services and tools for a new generation of machine learning and deep learning services using Python H20.ai.

8. Opta services, setup of SOLR search from collected Sports Data, and sports fixtures (uuids for sports venues, players, clubs, leagues, tournaments, seasons) .
9. Setup of Oracle Big Data Discovery visual tool over on-premissine BDA cluster.
    1. Usage of lake with published Apache Hive in Spark context jobs.
10. Setup of H2o Deeplearning and Boosting solutions, using R.
    1. Implementation of decision pipelines for content distribution for DAZN.
        1. A solution T-POC for Streaming AVGBitRate evaluation of normal/abnormal conditions, in http://bit.ly/2nFHYpf .
        2. A solution T-POC for Malware/Phishing intrusion detection and prediction. Designed with url based evaluation , based on external/internal data and conditions, in http://bit.ly/2qAISok .
11. Establishment of Fuel BigData .
    1. An entity dedicated for Ai consultancy.
    2. Follow site for more details https://fuelbigdata.com .

Project roles :

3. Lead Architect for the new Sports Media solution with third party tracking data , being matched with internal insights of sports event data. Production of new generation of Sports insights, from an event, to an even here, where, with this intervenients around the events. Control of the Software Development and Data science function support of new development algorithms and tools. Represented Perform Group at the Oracle OpenWorld 2017 in Intel Keynote sessions where solution was presented with or Partner Oracle (youtube link) . Presented the project in a variety of Oracle and Cloudera events (cloudera link).
4. Lead Architect for BigData solution for the new generation of Sports Fans insights across social media, designed and implementation of the social media collection tools, using Data streams from Omniture and from Google Analytics collection Google DFP for Adds Data, Gygia social registrations for Facebook, Twitter and G+. Implementation of Google Analytics, AdWord solution and Google Ads(ex- DPF) for market understanding and segmentation. Implementation of a new generation of Marketing Cloud DMP, for Goal.com and DAZN.com brands using Bluekai from Oracle Cloud. Full deployment and end to end with reporting using Oracle big Data , and Bluekai Cloud solutions.
5. Lead Architect for a Sport integrity for Betting solutions. Used Oracle Big Data SQL (3.1) for data servicing of a new generation of Spark/ML and SparkLyR context analysis over sports Data. Publication of data from Hive database , with transformations over collected data in SOLR search, for abnormal pattern detection.
6. Lead Architect for the new DAZN EDW solutions, coordinating the teams to scale a new platform for AWS Cloud with Big Data using solutions, AWS Glue, AWS RedShift/Spectrum, AWS EMR, AWs Kinesis.

- Senior Software Developer - EDM Big Data

(From January 2015 – October 2015)

SKY , CBS - EDM Osterley (**https://corporate.sky.com/contact-us**) (Permanent)

Business or sector: Systems Migration from Cloudera CDH 4.3 to Cloudera CDH 5.2 with ETL and CRM techniques for Media and Core product content Analytics.

Technologies used:

12. Designing ETL Solutions with Apache Spark 1.3 and production of insights for ElasticSearch 2.3.4 and Kibana 3.0.x.
13. Designing of ETL applications with Scalding and Cascading in Scala and using their core API for MapReduce 1.0 .
14. Design of new Flume agent ingestion processes, from near real-time data updates to Hadoop data sinks.
15. Big Data analytics with Cloudera CDH 4.3 and CDH5.2 in Hadoop ecosystem using Hadoop Streaming API for Scalding and with Cascading. Usage of Scoop connections for data ingestion from Oracle dataBases, Netezza and ABN Issue . Also implement analytics Jobs with scripting in PIG, HIVE and Impala.
16. ETL processing and BI analysis using Python UDF(s) in HIVE .
17. Restful API implementations, for data consumption and ingestion to data sinks. (FLUME with Plugging ingestion of HTTP consumers)

Project roles :

7. Development of a new solution for ETL on Mobile visualized content. Used JSON, CSV and XML ETL and parsing tools of Cascading and Scalding.
8. Prepare and migrate data products in between Cloudera CDH4.2 and Cloudera CDH5.2.
9. Daily support of production data and analytics.

- Software Consultant Big Data Analytics, Cloudera CDH

(From November 2014 – to December 2015)

Contractor from Gravitas for (QUDINI) (Contractor)

Business or sector: Systems design and migration from Cloudera CDH 4.3 to Cloudera CDH 5.2 with ETL techniques for Product content Analytics and CRM data. .

Technologies used:

18. Big Data analytics with Cloudera CDH_5.1 , Hadoop ecosystem using Scoop connections and doing analytics with scripting in PIG, HIVEQL and Impala.
19. ETL processing and BI analysis using Hive with Python UDF(s) and HIVEQL.
20. SOA implementation with J2EE framework, web Services in Struts and Spring API.

Project roles :

10. Design and implementation of a new Hadoop Cluster for a Startup company, specialized in software for Restaurant Queue Management in London. Design of the solution using Cloudera CDH5.1 and implementation of packages, Hadoop, Yarn, MapReduce 2, Hive, Impala, Sqoop, HBase.
11. Project not finished due to issues in Startup.

- Software Systems Data Analyst and Senior Software Developer / Analyst

(From May 2010 – to October 2014)

Fidelity Global, EMEA Watford (**www.fisglobal.com/EMEA/UK**) (Contractor)

ImpactZero Software (**www.impactzero.pt**) (Senior Software Analyst)

Business or sector: Systems Migration with ETL and CRM techniques for Core Banking, customer data and transactional data from payment systems. New data transformation processes for funds in-clearing and migration payment redirections in domestic and international schemas as FPS, BACS, CHAPS, SEPA and following standard formats SWIFT MT103 , BACS AUDIS and BACS ADDACS DDI. Product development and life cycle support for Business and Current account product.

Technologies used:

21. Big Data analytics with MySQL, and MongoDB ecosystems and more recently with Hadoop ecosystem using Scoop connections and doing analytics with scripting in PIG, HIVEQL and Impala.
22. ETL processing and BI analysis using Hive with Python UDF(s) and HIVEQL .
23. SOA implementation with J2EE framework, web Services in Struts and Spring API.

Project roles :

12. Systems and Data Reporting Control Analyst in projects for Barclays Direct formerly known ING Direct UK. Data Quality check for Data WareHousing, Business Intelligence and Customer Relationship management.
13. MDM Analysis, Design of processes on business requirements with technical implementations for Data Migration in Core Banking applications. Active participation in integration discussions and cross-functional issues resolution control. Active participation in all the different stages of the migration project: requirement gathering, analysis, build, implementation, system testing and performance control. Participation on the go-live and coverage of data migration.
14. Technical specialties: 4+ years of experience in Banking Software Solutions, performed different roles and tasks, from Functional, Technical Analysis, Build and Implementation. Covered several domains, such as Core Banking CRM, B2C/B2B, Web Services, SOAP, SOA, Transactional Data Capture and Historical log processing. Use RPC techniques and technical solutions as :
    1. Implement xml parsers for ISO 20022 message standard and SEPA payment exchanges.
    2. Develop xml parsers for real-time customer SMS, using connector IBM WebSphere MQ V7.0 client components.
    3. Develop new core product processes using Web services with Struts, and Spring API for Customer bank applications using SOAP MEP patterns.
    4. Develop new business account product in FIS Profile Core banking : Front-end in Java (Web Services or SOA) and back-end in GT-M VM, scripting in KSH for AIX console.

-

Software Analyst, Developer and Tester

(From June 2008 – to April 2010)

T-Systems Iberia Barcelona / D-Core Lisbon (**www.t-systems.pt/**) .

Started as a junior software analyst and developer.

Worked in SLI (Sales system) applications for VW UK and VAESA.

Business or sector : Analysis, Development and Software Testing using ETL techniques for car sales systems.

Technologies used:

1. ETL Technologies in mainframe zOS ecosystem using COBOL, JCL and CICS.

Project roles :

1. Actively participate in the different stages of the project : analysis, build, testing and implementation.
2. Actively participate in new ETL processes for Migration in between Mainframe systems (From VAESA in Spain to VW UK in Wolfsburg).
3. Development of new BI and SLI logic customised to RDD of UK customer .
4. Use technologies for process control and Data : JIRA and HP Quality Centre.
5. Use technologies for data ETL: Mainframe host with COBOL, JCL, CICS and DB2 database.
6. Development and testing control of SLI BI using Java Swing interfaces.

EDUCATION AND TRAINING

- Cloudera Data Science Training . London UK : (2016 -)

1. The fundamentals of Cloudera Big Data Lake.
2. Fundamentals of Data Science with Python, and R.
3. Practical use case analytics using ALS recommendation model.
4. Comparative analysis of models using R plotting, via ggplot2 and rCharts.

- Cloudera Administration Training. London UK : (2016 -)

5. The fundamentals of Cloudera Big Data Lake.
6. Setup of a 5 Nodes Cluster using AWS and EMR.
7. Scale of a Cluster and setup of Services in a Cluster.
8. Root cause analysis for issues and problems in a cluster.
9. Scalability in a CDH cluster, adjust Dynamic Pool of resources and Static Pool of resources.

- Cloudera Data Analyst Training. London UK : (2014 -)

10. The fundamentals of Apache Hadoop and data ETL (extract, transform, load), ingestion, and processing with Hadoop MapReduce and tools
11. Organizing data into tables, performing transformations, and simplifying complex queries with Hive
12. Joining multiple data sets and analysing disparate data with Pig
13. Performing real-time interactive analyses on massive data sets stored in HDFS or HBase using SQL with Impala.
14. Connectivity of Hadoop HDFS and Oracle 11g or MySQL using Sqoop.

- Continuous Training in FIS Profile Technologies. FIS – Fidelity Global : (2010 - 2014)

1. Training in Core Banking B.I. - Data ETL and CRM . Usage of Falcon Fico , Experian extraction and IBM Cognos.
2. Training in Core Banking technical solutions from FIS Global.
3. Currently also doing training for junior resources on-site and off-shore.

- Training in z-OS Mainframe Technologies.T-Systems - IBM Portugal : (2008 - 2009)

Training for integration in T-Systems project. IBM Mainframe z/OS modules :

- ES10 – Fundamental System Skills in Z/OS.
- AD40 – COBOL Programming Fundamentals.
- CF82 – DB2 Programming Workshop for Z/OS.
- CI01 – CICS Fundamentals.
- CI17 – CICS Application Programming I.
- CF96 – DB2 for z/OS Application Performance and Tuning.

- Studies in Mathematics and Computer Science

Universidade Do Minho, Braga Portugal : (2002 - 2007)

Studies this Mathematics and Computer Science 5 year plan. An pre-Bologna Education Agreement plan with Studies in Mathematics and Advanced Software Engineering. It included advanced imperative and functional language paradigms, knowledge representation, artificial intelligence and modern distributed systems architectures. During this period have learned to work in Unix systems and software development in technologies JAVA6, Scala, Python, R, Bash, Perl, Erlang, Haskell, C, C++ . The program of degree was highly oriented to project development from requirement gathering and analysis to implementation and product implementation and ownership.

PERSONAL SKILLS

Mother tongue(s) Portuguese

Other language(s)

	UNDERSTANDING		SPEAKING		WRITING
	Listening	Reading	Spoken interaction	Spoken production	
English	C2	C2	C2	C2	C2
	Certificate				
Spanish	B2	B2	B2	B2	B1
	Certificate				

Levels: A1/2: Basic user - B1/2: Independent user - C1/2 Proficient user
Common European Framework of Reference for Languages

Communication skills

Good communication skills gained through my experience as a resource on customer and representing the company in requirement gathering meetings, major problem resolution meetings and in continuous support meetings.

Organisational / managerial skills

Leadership positions :

- For Perform Media Group (now DAZN Group) was the Head of BigData. I did mentoring of teams in the space

of Big Data technologies, conducted team knowledge refresh sessions and conducted the adoption of Big Data frameworks like Spark, managed near-shore teams in Poland (Katowice) and in Slovakia (Kosice).
- For FISGlobal / ING Direct did team integration via phone and video conference calls, as currently responsible for a small team of 5 people in between on-site (UK Reading) and off-shore resources (India Bangalore).

Job-related skills

Highly skilled in software development life cycle, for Media systems in the big data stack also before in the maintenance, support, builds and packaging of core banking software solutions (responsible in FIS team for software life cycle maintenance and production deployment, including migration and services closure).

Computer skills

Full domain of JIRA Agile Project management tool. Full domain of test management HP Quality Center. Full domain of Microsoft Office tools as Visio, Excel, Word and Project. Software development with in SOA, Java Web Services, MySQL, PostgreSQL, banking product development. Also knowledge in scripting for UNIX / AIX and Z/OS systems using Bash, KSH and JLC.

Other skills

Storytelling skills / Soft skills, capable to communicate ideas and manage the art of communication in holistic human side. Worked in the family business (in accountancy and stock management areas). Cyclist on free weekends in both road and mountain, also a daily office commute when cycling is possible.

Driving licence

Driving Licence, category B

NobleProg
The World's Local Training Provider

course completion

CERTIFICATE

has been presented to

Joao Cerqeuira

for the completion of

Domain Driven Design - bespoke

Trainer
Antonio Radesca

Venue
London

Duration
1 Day

date of issue:
2019-06-24

e-certificate:
https://cert.nobleprog.com/authenticate
Certificate ID: 586523
Authentication Code: cde51

Oracle Cloud Customer Connect

Certificate of Achievement

Presented to

Joao Cerqueira

Congratulations! Through your contributions to the community, you have earned the Green Ribbon badge.
We truly appreciate you sharing your knowledge and expertise with the community,
and we look forward to seeing even more of your valued posts!

January 01, 2019

ORACLE
Cloud Customer Connect

Jeff Caldwell
Community Director

CLOUD CONNECT

cloudera®

Certificate of Attendance

is hereby granted to

Joao Cerqueira

To verify that he/she has attended

Data Science at Scale using Spark and Hadoop

Sarah Sproehnle
VP, Educational Services

Cloudera, Inc.
www.cloudera.com

October 5th 2016
Course Date

cloudera®

Certificate of Attendance

is hereby granted to

Joao Cerqueria

To verify that he/she has attended

Cloudera Data Analyst Training: Using Pig, Hive, and Impala with Hadoop

VP, Educational Services

Cloudera, Inc.
www.cloudera.com

August 12-14th
9AM-5PM
Date

Certificado

A Companhia IBM Portuguesa, declara, para os devidos efeitos, que:

João Pedro Afonso Cerqueira

frequentou o(s) seguinte(s) Curso(s):

CÓDIGO	DESCRIÇÃO	INÍCIO	FIM
ES10	Fundamental System Skills in z/OS	03-07-2008	09-07-2008
AD40	COBOL Programing Fundamentals	10-07-2008	15-07-2008
CF82	DB2 Programming Workshop for z/OS	16-07-2008	18-07-2008
CI01	CICS Fundamentals	21-07-2008	22-07-2008
CI17	CICS Application Programming I	23-07-2008	28-07-2008
CF96	DB2 for z/OS Application Performance and Tuning	29-07-2008	01-08-2008

Os cursos decorrem no horário das 9:30 às 12:30 e das 14:00 às 17:30, em dias úteis.

Carlos Ved
IBM Learning BTO Delivery Leader

IT Education Services Portugal

Universidade do Minho
Serviços Académicos

Carla Isabel Pereira Lavrador, Registrar of the University of Minho, hereby certifies, according to University records, that João Pedro Afonso Cerqueira, born in the parish of Caldelas, in the county of Amares, in the district of Braga, son of Joaquim Jorge Mota Cerqueira and of Maria da Conceição Silva Afonso Cerqueira, obtained the following passing grades in the "Licenciatura" in Computer Science:

Course Unit	Regime (1)	Type (2)	Academic Year	PT Mark (3)	ECTS Mark (4)	Credits (5)	Obs. (6)(7) (8)
Linear Algebra	1	Ob	2002/2003	10	E	5.0	
Calculus	1	Ob	2002/2003	10	E	5.0	
Discrete Mathematics I	1	Ob	2002/2003	10	E	6.0	
Discrete Mathematics II	2	Ob	2002/2003	11	D	3.5	
Programming Paradigms II	2	Ob	2002/2003	10	E	3.5	
Programming Paradigms I	1	Ob	2003/2004	12	C	3.5	
Analysis	2	Ob	2003/2004	11	D	5.0	
Linear Algebra and Analytical Geometry	2	Ob	2003/2004	13	C	8.0	
Programming Paradigms III	1	Ob	2003/2004	10	E	3.5	
Automata and Turing Machines	2	Ob	2003/2004	11	D	3.5	
Computational Mathematics	2	Ob	2003/2004	10	E	4.0	
Programming Paradigms IV	2	Ob	2003/2004	12	C	3.5	
Topics in Analysis	1	Ob	2004/2005	10	E	4.0	
Computer Architecture	2	Ob	2005/2006	13	C	3.5	
Numerical Analysis II	2	Ob	2005/2006	10	E	6.0	

(1) – Regime: A – Annual 1 – 1st Semester 2 – 2nd Semester
(2) – Type: Ob – Compulsory Op – Optional
(3) – Grade: On a 0 to 20 grading scheme, a minimum grade of 10 in each course unit is necessary to pass.
(4) – ECTS Grading Scale

ECTS Grading Scale	% of approved students
A	10
B	25
C	30
D	25
E	10

In order to establish the course units grades for each group of the ECTS grading scale, the distribution of students with passes in the previous five academic years and/or in a total of at least thirty students is taken into consideration. When a cohort of this scale is impossible, the ECTS grading scale is replaced by the use of an institutional conversion table based on all course units grades of all University of Minho's Degree Courses from the previous five academic years. It is important to notice that not all groups of the ECTS grading scale might be represented.

(5) – ECTS Credits: 1 academic year= 60 1 semester = 30 1 term = 20
(6) – Course units subject to academic recognition
(7) – Extracurricular course units
(8) – Curricular units accredited by demonstration of professional competence and/or other training

This same certificate carries the embossed seal of this University.

The Registry of the University of Minho, on the 25th of July, 2014.

The Registrar,