Crecemas: A transactional data-based big data solution to support a bank's corporate clients with their commercial decisions

Antonio Martín Cachuán Alipázaga Willy Alexis Peña Vilca
Big Data Center of Excellence
Banco de Crédito del Perú
Lima, Perú
{acahuan,wpena}@bcp.com.pe

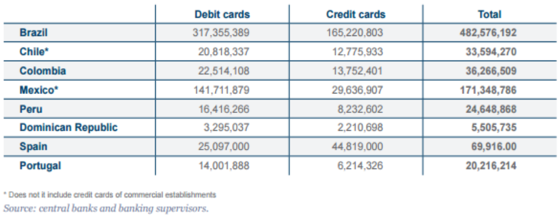
Abstract

Peru is recognized in the region of Latin America as a country of entrepreneurs so enterprises that have grown rapidly are more commonly found in the market; however a great number of them cease their operations in a short time due to lack of capabilities to use data to know better their clients and their competition. As a solution to help this kind of business to be sustainable in time, exists new ways to process and analyse data generated by clients through the transactions they made with their credit or debit cards. In addition, this information is usually manage by banks and financial services which with the help of new technologies as Big Data and Cloud Computing, that banks use in the daily basis, can help their corporate clients to achieve their goals by providing them with aggregated information through analytic indicators about their clients and competition.

In this article, we show the Big Data architecture of the web platform "Crecemas", which was developed in 16 weeks under agile methodologies with the Scrum framework and using Cloud Computing technologies. In this web, KPIs are shown using anonymized transactions about the company, its clients and its competitors, which is helpful to support commercial decisions. Currently the platform handles 200gb of information with 7 worker nodes and 3 master nodes and is used by almost 500 different companies in Peru.

1 Introduction

The search for dynamism the generation of work in a country is mainly to support new entrepreneurs in order to contribute directly to innovation (Harman Andrea, 2012). All this improves the country's economy by bringing welfare to its inhabitants. The region of Latin America and the Caribbean is characterized for the entrepreneurship, Peru is not the exception. It is a country which is occupying the eighth place in a group of 60 economies according to the Global Entrepreneurship Monitor, but it also has the highest rate in failure (Donna Kelley Slavica Singer, 2016), one reason for this result, is the low index of strategic alliances (Global Innovation index, 2015), which means that large companies are not actively seeking to do business with smaller companies, all this has an impact on the competitiveness index of the country where it is ranked 69th out of 140 countries (Donna Kelley Slavica Singer, 2016). In addition, most of them do not take advantage of the information they generate in the daily basis , because they don't have access to all the data that they need to accomplish this or don't have the required capabilities inside the company to bring strategic insights (Brynjolfsson et al., 2011).



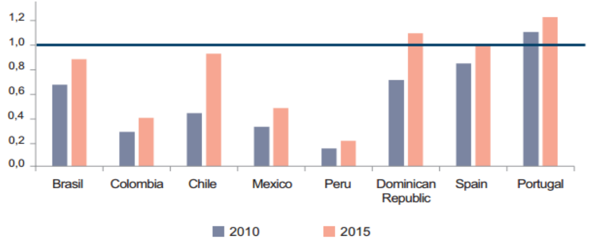
	Debit cards	Credit cards	Total
Brazil	317,355,389	165,220,803	482,576,192
Chile*	20,818,337	12,775,933	33,594,270
Colombia	22,514,108	13,752,401	36,266,509
Mexico*	141,711,879	29,636,907	171,348,786
Peru	16,416,266	8,232,602	24,648,868
Dominican Republic	3,295,037	2,210,698	5,505,735
Spain	25,097,000	44,819,000	69,916.00
Portugal	14,001,888	6,214,326	20,216,214

* Does not it include credit cards of commercial establishments

Source: central banks and banking supervisors.

Figure 1: Number of credit and debit cards in circulation (2015)

On the other hand, according to the Tecnocom report of 2016 (Tecnocom, 2016), the annual average of transactions with debit and credit card are 15 and 5 respectively. That happens because the number of debit and credit cards in circulation in Peru is growing steadily over the last five years, as is shown in Fig. 1, with a growth of 9.6% in debit cards. This translates into an increase in the use of payment cards usage. As you can see in Fig. 2, Peru had a ratio of POS card vs. Cash withdrawals of ATM of less than 0.3, which means that there is a great potential for growth.



Source: central banks and banking supervisors.

Figure 2: Ratio of value of POS card transactions to ATM withdrawals (2010-2015)

The digital transformation is a chance for the organizations to get better at managing information as a strategic asset and Big Data is a game changer who adds more sources to the mix. However, unless the lessons of the productivity paradox are applied (E. Brynjolfsson, 1994), these changes will only serve as distractions. Companies that anticipate the changing needs of the volatile marketplace and successfully implement new technologies, place themselves in a great position to overcome their competitors (Earley, 2014).

Consequently, there are favorable circumstances for banks and financial services companies interested in Digital Transformation who are willing to use the information generated by the users of credit and debit card to help the smaller and newer companies to grow in the country's economy. To accomplish this, banks can provide aggregated information obtained from sales transactions so that the business can make better decisions.

2 Literature Review

2.1 Concepts

2.1.1 Apache Hadoop

Hadoop is an open-source software for reliable, scalable and distributed computing. It brings a framework that allows the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. The project includes these modules (The Apache Software Foundation, 2007a)

- Hadoop Common: The common utilities that support the other Hadoop modules.
- Hadoop Distributed File System (HDFS): A distributed file system that provides high-throughput access to application data.
- Hadoop YARN: A framework for job scheduling and cluster resource management.
- Hadoop MapReduce: A YARN-based system for parallel processing of large data sets.

This distributed framework has been adopted by different vendors, such as Cloudera and Hortonworks who have added important features, such as data governance and security compliance, which give to this powerful technology an enterprise attractive characteristic. In addition, the community has played an important role; for example, some of the main tools included by the Apache Foundation are the following:

- Apache Hive: data warehouse software that facilitates reading, writing and managing large datasets residing in distributed storage using SQL (The Apache Software Foundation, 2011).
- Apache HBase: Hadoop database that brings the possibility of random accesses and real-time reading and writing to Big Data storages (The Apache Software Foundation, 2007b).

2.1.2 Cloud Services

Cloud Services are applications or services offered by means of cloud computing. Nowadays, nearly all large software companies, such as Google, Microsoft and Amazon, are providing this kind of service. In addition, cloud computing has revolutionized the standard model of service provisioning allowing delivery over the Internet of virtualized services that can scale up and down in terms of processing power and storage.Cloud computing also provides strong storage, computation, and distributed capabilities to support Big Data processing. In order to achieve the full potential of

Big Data, it is required to adopt both new data analytics algorithms and new approaches to handle the dramatic data growth. As a result, one of the underlying advantages of deploying services on the cloud is the economy of scale. By using the cloud infrastructure, a service provider can offer better, cheaper, and more reliable services. Cloud services offer the following schemas of services (Campbell et al., 2016).:

- SaaS: Costumers do not have control over the hardware and software level configurations of the consumed service.
- PaaS: Platform usually includes frameworks, developing and testing tools, configuration management, and abstraction of hardware level resources
- IaaS: Costumers can hire a hardware-level resources.
- DBaaS: Database installation, maintenance and accessibility interface are provided by a database service provider.

2.1.3 Agile Methods

Agile methods are contemporary software engineering approaches based on teamwork, customer collaboration, iterative development, and constantly changing people, process and tech. This approach diverts from traditional methods which are software engineering approaches based on highly structured project plans, exhaustive documentation, and extremely rigid processes designed to minimize change. Agile methods are a de-evolution of management thought predating the industrial revolution and use craft industry principles like artisans creating made-to-order items for individual customers. Traditional methods represent the amalgamation of management thought over the last century and use scientific management principles such as efficient production of items for mass markets. Agile methods are new product development processes that have the ability to bring innovative products to market quickly and inexpensively on complex projects with ill-defined requirements. Traditional methods resemble manufacturing processes that have the ability to economically and efficiently produce high quality products on projects with stable and well-defined requirements (Trovati et al., 2016).

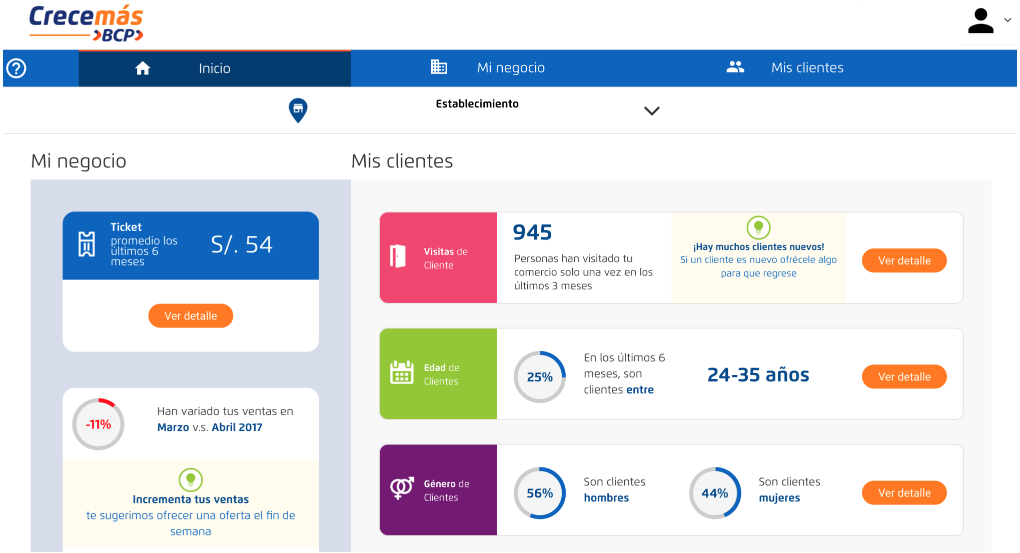
2.2 Related Work

Nowadays, data is being generated at an unprecedented scale. Decisions that previously were based on guesswork or handcrafted models of reality, can now be made using data-driven mathematical models. However, this increase of the amount of data and the variety of formats has put new challenges on the table; for example, is more complicated to deal with this kind of data and have a high performance process with the technology that many organizations are using in the daily basis. For that reason, Big Data has the potential to revolutionize much more than just research the batch processing of data, this technology has come to enable analysis on every aspect of mobile services, society, retail, manufacturing, financial services, life science and others (Jagadish et al., 2014). In addition, in order to accomplish the successful use of this kind technology, organizations need to have an enterprise infrastructure that could support this initiative with the goal of maintain and run the transformation process in an efficient way. That said, purchasing and deploying equipment in a short term is important, in order to reduce the delivery time of the solution, Cloud Computing is a revolutionary mechanism that is changing the way that enterprise enable hardware and software design and procurements in an efficient and economical way; with this in mind, the possibility to enable an infrastructure in more flexible environments such as those of the cloud, makes the use of this type of technology much more attractive, in order to provide end users with fast and useful results for them (Philip Chen and Zhang, 2014).

With all these great benefits the use of Big Data technologies and Cloud Computing are a perfect combination to start this journey. Also, as mentioned in (Vurukonda and Rao, 2016), it is important to keep in mind that although the cloud is an attractive option, the biggest challenge regarding cloud is the security and regulatory issues about a company's customer data in such environments, so it also carries great challenges that are been currently working (Hashem et al., 2014).

3 Proposed Solution

The solution is structured in three stages that range from obtaining the data directly from the internal



Mi negocio

Ticket promedio los últimos 6 meses S/. 54

-11% Han variado tus ventas en Marzo v.s. Abril 2017

Incrementa tus ventas
te sugerimos ofrecer una oferta el fin de semana

Mis clientes

Visitas de Cliente
945
Personas han visitado tu comercio solo una vez en los últimos 3 meses

¡Hay muchos clientes nuevos!
Si un cliente es nuevo ofrécele algo para que regrese

Edad de Clientes
25% En los últimos 6 meses, son clientes entre 24-35 años

Género de Clientes
56% Son clientes hombres
44% Son clientes mujeres

Figure 3: Dashboard from crecemas.com

sources, loading them to the cloud and transforming them to be able to calculate the indicators ending in their visualization in the web (Fig.4).

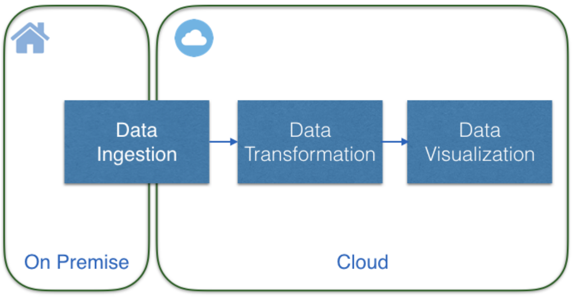


Figure 4: Proposed solution diagram

The entire solution for Crecemas http://www.crecemasbcp.com/ (Fig. 3) was developed in 16 weeks led by a scrum master and a total of 13 people dedicated exclusively where each member was grouped according to one of the three roles (Table 1).

- Business: Dedicated to engage internal areas and avoid possible business stoppers. In addition the design of the KPIs.
- Data: Responsible for Big Data stage.
- Development: In charge of the visualization of the data and the web.

3.1 Data Ingestion

For this stage, different information extraction processes were built in order to obtain the information from the Enterprise Data Warehouse and then store it into a file server. This processes also perform some field filtering and record according to the requirements for the KPI construction. Regarding the regulatory constraints, the main objective of this processes was to tokenize some sensitive fields that couldn't be stored in a Cloud environment (e.g., client's names, client's address, card number).

In addition, each file generated has a control file to perform a validation in the data upload process. This file contains the number of exported records and the date in which the file was processed. For this reason, data ingestion worked with two files: the first one just with credit card's transactions data (with extension *.dat*) and the other one just with control data (with extension *.ctrl*)

The files that were extracted with these processes have the following types:

- Daily master tables: That are completely loaded.
- Daily incremental tables: That have information of one day and are stored in order to have history of the data.
- Monthly master tables: That are incrementally loaded.

Business	Data	Development
1 Product Owner	1 Big Data Architect	1 Back-end developer
1 Navigator	2 Data Engineer	2 Front-end developer
1 Research	2 Data Expert	1 UI Expert
		1 UX Expert

Table 1: Crecemas team

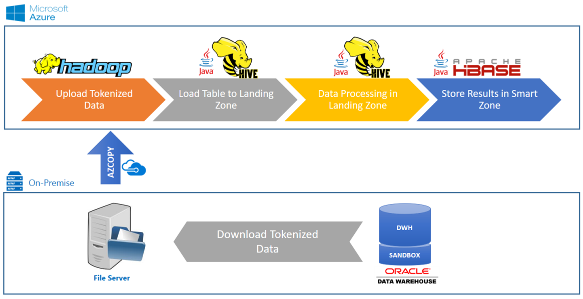


Figure 5: Solution proposed for data ingestion

As can be seen in Fig. 5, The ingestion process as performed in the following way: All the orchestration of the processes in the On-Premise environment was made by the enterprise scheduling tool, which controls and executes the information extraction processes. Once this processes have finished running, one final job executes an AzCopy command (Multi-thread tool to upload data to Microsoft Azure cloud environment), which is in charge of uploading aforementioned files from the file server to the Linux servers(that were created to deploy the Big Data technology) in the cloud environment. Next, in these servers another job is executed which invokes a Ad-Hoc client that uploads the data from Linux to HDFS, to finally upload the created Apache Hive tables, which are used for the KPI's construction. This client (Apache Hadoop and Apache Hive) was created using the Maven Artifacts from Cloudera, who is the Big Data provider selected for this project.

3.2 Data Transformation

After the data ingestion process, data is stored in the Hadoop ecosystem (HDFS) within a Clouderas cluster which uses 7 worker nodes and 3 name nodes (1 main and 2 for backup) its architecture can be observed in Fig. 6.

3.2.1 Landing Area

This is the initial zone where the data in HDFS is located as they were loaded by the process of data ingestion. Data can be accessed through a database composed of External Tables created using HQL (Hive Query Language). First, inconsistencies in the volume of information processed by each table and business-level inconsistencies in the values are reviewed (birth date, sex, foreign characters and incongruent transactions). Then a data cleaning process occurs, which eliminates duplicates and replaces empty nulls. Finally, the data moves to a new HDFS location and is created a Hive database called Tmp Transformation Area.

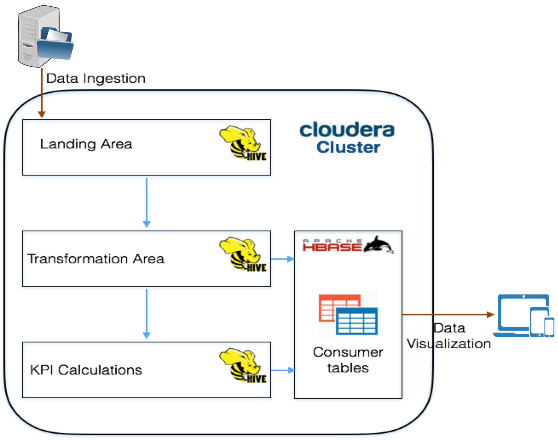


Figure 6: Data transformation process

3.2.2 Transformation Area

This area consists of two databases in Hive: the first one (Tmp Area Transformation) contains the tables generated by the previous component with potential update, adds, eliminations if necessary. The second database is the result of a transformation process of the first DB and consists of a 'Tablon' (a large table with more than 100 columns) that consolidates all the information at the level of transactions and commerce including tokenized customer information (such as sex, age, date, educational level and economic level). This large table is used for the KPI Calculations component.

3.2.3 KPI Calculations

In this area, the 'Tablon' is used to generate 7 tables, each linked to one or more KPIs that are detailed in the Data Visualization part. The tables are in a Hive database, which is connected to NoSQL tables in HBase responsible to display the reports on the web.

3.2.4 Consumer Tables

There are 7 tables of type <*Key, Value*> that are consumed using Java applying a Facade Design Pattern (details in the section 3.3).

Today we handle almost 200 gb of historical data and the entire transformation process is executed daily in 1 hour for a volume of information of approximately 12gb (11gb for master tables like customers and business and 1gb for transactions) which accumulate during the month. Also every month there is a process that goes through all the stages and takes about 1 hour with a volume of 10 gb of data (mainly other master tables related to business location, local geography).

At the steady state, daily processing should be 20gb and monthly processing 15gb which means a total of 600gb historically[1] (Table 2)

	Daily	Monthly	Historic
Actual	12gb	10gb	200gb
Projected	20gb	15gb	600gb

Table 2: Data size processed

3.3 Data Vizualization

Data Visualization solution for this project was a web platform. (Figure 3). For that reason, this stage is a real-time request processor that allows to query the Apache HBase database, in order to obtain the data and show the results to the end user.

The workflow for this stage is shown in Fig. 7.

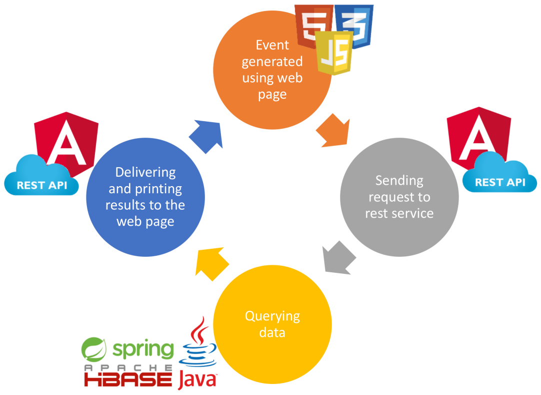

Event generated using web page
Sending request to rest service
Querying data
Delivering and printing results to the web page
REST API
REST API
spring
APACHE HBASE
Java

Figure 7: Data Visualization Diagram

This stage is composed by the following components:

- External Communication: Services that allow consumption of the information in the Big Data environment.
- Client Communication: In charge of establishing the remote connection with the application back-end from web pages.
- Graphics: Statistics graphs on the web pages.
- Web Page: Represent all the system's web pages.
- Reporting: Allows to access to the system repositories in order to perform analysis and create new reports.

4 Conclusion and Future Works

The proposed solution allows enterprises to enable Big Data capabilities inside the organization making more efficient batch processes and reducing solutions time to market. In addition, the use of Cloud environments facilitates the adoption of technologies that require an intensive infrastructure deployment. Likewise, the agile framework used by the project, demonstrates that having the client in the center of all decisions and solutions allows a product to be created in a short time and with great value to the clients.

For future works, the process developed in Apache Hive could be migrated to different technologies that improves the processing speed. For example, Apache Impala (Kornacker et al., 2015) or Apache Spark (Zaharia et al., 2016) are great

[1] Historic data from last 2 years.

options, because this technologies offer a different engine of execution that brings new capabilities to the solution proposed. On the other hand, the information that users are generating inside the web could help to make important improvements in how the company knows better their clients in order to offer solutions that could help to accomplish their main goals.

References

Erik Brynjolfsson, Lorin M. Hitt, and Heekyung Hellen Kim. 2011. Strength in Numbers: How does data-driven decision-making affect firm performance? *ICIS 2011 Proceedings* page 18. https://doi.org/10.2139/ssrn.1819486.

Jennifer Campbell, Stan Kurkovsky, Chun Wai Liew, and Anya Tafliovich. 2016. Scrum and Agile Methods in Software Engineering Courses. In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*. ACM, New York, NY, USA, SIGCSE '16, pages 319–320. https://doi.org/10.1145/2839509.2844664.

Mike Herrington Donna Kelley Slavica Singer. 2016. Global Enterpreneurship Monitor. http://www.gemconsortium.org/report/49480.

E. Brynjolfsson. 1994. The Productivity Paradox of Information Technology: Review and Assessment Center for Coordination Science. http://ccs.mit.edu/papers/CCSWP130/ccswp130.html.

S. Earley. 2014. The digital transformation: Staying competitive. *IT Professional* 16(2):58–60. https://doi.org/10.1109/MITP.2014.24.

Harman Andrea. 2012. *Un estudio de los factories de exito y fracaso en emprendedores de un programa de incubacion de empresas: Caso del proyecto RAMP Perú*. Master's thesis, Pontificia Universidad Catolica del Peru.

Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, and Samee Ullah Khan. 2014. The rise of Big Data on cloud computing: Review and open research issues. *Information Systems* 47:98–115. https://doi.org/10.1016/j.is.2014.07.006.

H. V. Jagadish, Johannes Gehrke, Alexandros Labrinidis, Yannis Papakonstantinou, Jignesh M. Patel, Raghu Ramakrishnan, and Cyrus Shahabi. 2014. Big data and its technical challenges. *Communications of the ACM* 57(7):86–94. https://doi.org/10.1145/2611567.

Marcel Kornacker, Alexander Behm, Victor Bittorf, Taras Bobrovytsky, Casey Ching, Alan Choi, Justin Erickson, Martin Grund, Daniel Hecht, Matthew Jacobs, Ishaan Joshi, Lenni Kuff, Dileep Kumar, Alex Leblang, Nong Li, Ippokratis Pandis, Henry Robinson, David Rorke, Silvius Rus, John Russell, Dimitris Tsirogiannis, Skye Wanderman-Milne, and Michael Yoder. 2015. Impala: A Modern, Open-Source SQL Engine for Hadoop. *Cidr* http://impala.io/.

C. L. Philip Chen and Chun Yang Zhang. 2014. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences* 275:314–347. https://doi.org/10.1016/j.ins.2014.01.015.

Tecnocom. 2016. Informe Tecnocom Tendencias en Medios de Pago. https://goo.gl/95th2L.

The Apache Software Foundation. 2007a. Apache Hadoop. http://hadoop.apache.org.

The Apache Software Foundation. 2007b. Apache HBase. https://hbase.apache.org.

The Apache Software Foundation. 2011. Apache Hive. https://doi.org/10.1002/ciuz.201500721.

Marcello Trovati, Richard Hill, Ashiq Anjum, Shao Ying Zhu, and Lu Liu. 2016. *Big-Data Analytics and Cloud Computing: Theory, Algorithms and Applications*. Springer.

Naresh Vurukonda and B. Thirumala Rao. 2016. A Study on Data Storage Security Issues in Cloud Computing. In *Procedia Computer Science*. volume 92, pages 128–135. https://doi.org/10.1016/j.procs.2016.07.335.

Matei Zaharia, Michael J. Franklin, Ali Ghodsi, Joseph Gonzalez, Scott Shenker, Ion Stoica, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, and Shivaram Venkataraman. 2016. Apache Spark: a unified engine for big data processing. *Communications of the ACM* 59(11):56–65. https://doi.org/10.1145/2934664.