The World of Data in 2019

As (would be) seen from my eyes

Will go through the value chain for data.

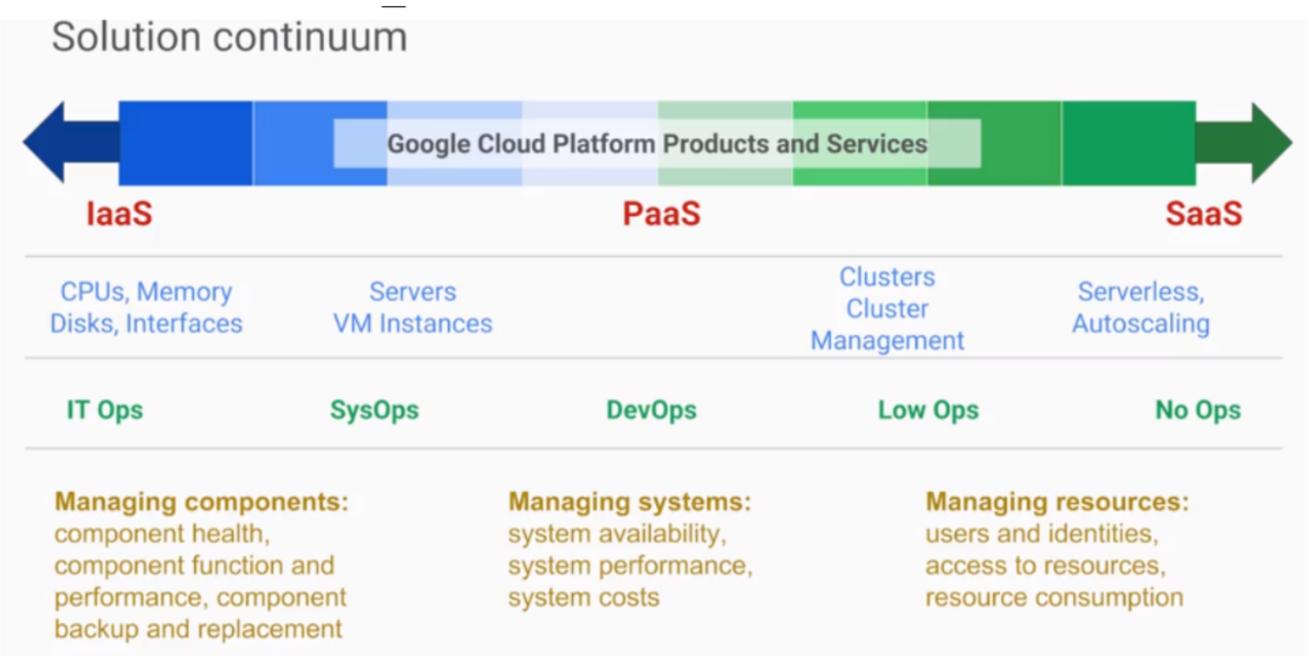
Working on this, it came to my mind that companies like Devoteam are to the business what cloud providers are for the infrastructure and IT teams. Drawing from this analogy the need would be more on the PaaS (Platform as a Service) side.

The Hadoop distributions world seems to have fallen apart. The complexity, difficulty of integration, lack of critical data governance are some of the reasons. Hadoop will continue to play some role for on-premise solution, legacy in particular, but with the current trend I do not see potential for growth.

Infrastructure

As we already discussed the trend to shift the focus from on-premise to cloud will continue. Further the interest on IaaS will go down as PaaS and SaaS will be stable and growing. Respectively the need for low level skills would be low - system administration and operations, software development. Knowledge and experience in popular and available platforms will be stable and growing.

(From GCP Course)



Solution continuum

Google Cloud Platform Products and Services
IaaS
PaaS
SaaS
CPUs, Memory Disks, Interfaces
Servers VM Instances
Clusters Cluster Management
Serverless, Autoscaling
IT Ops
SysOps
DevOps
Low Ops
No Ops

Managing components: component health, component function and performance, component backup and replacement

Managing systems: system availability, system performance, system costs

Managing resources: users and identities, access to resources, resource consumption

Storage

Each cloud provider uses abstraction layer to isolate the actual storage details from applications. For that reason I do not find the storage technologies and topics relevant, except for understanding high availability, redundancy and disaster recovery principles. (Available as options from cloud providers)

Data Integration and Data Management

Currently there is some gap with end to end data integration and data governance. This makes me believe that Data Integration Tools as Informatica, Talend, IBM Information Server will continue to play important role and will probably get back attention.

Taking in account the need for cloud integration - Microsoft - SSIS (SQL Server Integration Services) and ADF (Azure Data Factory), Snowflake.

NoSQL databases, e.g. Key-value, although interesting, didn't prove value for Data Integration and Management.

The role of IoT, streaming data, real-time processing and analytics leads to:

- Kafka + Kafka Connect
- Elastic Search
- NiFi (and the whole Hortonworks HDF stack)

Professionals, working in Data Management need to be portable. The need for cross platform skills portability leads to:

- SQL - very strong skills are must - beyond simple select
- Good understanding of the parallel processing architectures and challenges is must. Experience with MPP systems, e.g. Teradata, DataStage etc. - nice to have
- Scripting - Linux Shell - ability to perform data management at the command line - bash, awk, Python - strong, must
- Data Modeling
    - understanding of different logical data models - challenges and applicability
    - understanding of physical formats - JSON, XML, CSV, Avro, Parquet, ORC
- Knowledge of languages like Scala, Java are not relevant as application programming applied to Data Integration proved to be counter-productive.
- Knowledge of Spark is good, although ability to apply SparkSQL in real life is of actual value.

Data Governance

IBM are heavily investing into Apache Atlas as general purpose framework. In my opinion it is far from maturity and is just a framework. Actual integration into platforms is not mature. Other major players, e.g. IBM Information Governance Catalog and Informatica cannot provide proper integration for non-proprietary platforms. To me this is a grey area which needs some attention.

In the context of Data Governance we could place GDPR. We could benefit from our experience in LG and potentially look for partnering with companies like Privacera, SecuPi, Privatar. This is still immature area which have to build architectural patterns. Given that it is regulatory required we can assume constant growth of interest.

Data Analytics

Still hasn't proved general public value. To me this will continue to be "in-house". The need for external Data Scientists will be minor. Over the time patterns will start to emerge and it might become commodity and require external consultants.

Business Intelligence and Visualization

- Kibana (around Elastic Search)
- Tableau
- Qlik
- MS Power BI

Solution Architecture and Design

Putting all the pieces together is important and often underestimated.

Marketing could probably help with identifying trends on the problems that the business strives to solve. Together we could build up a catalog of solutions for typical business problems.

Summary

From team perspective, strong knowledge and understanding of general Data Management and Data Integration techniques and principles are critical for success. Strong knowledge of the integrated stack of at least one major cloud provider. Broad knowledge and skills with focus on cross platform, portable skills, e.g. SQL, Linux, etc.

From tool perspective:

- Informatica, Talend, Snowflake, MS Azure Data Factory, Kafka, Elastic (with Kibana), NiFi (HDF)