BIG DATA

5 - Conclusion

IS HADOOP STILL A THING?

WHAT 'KILLED' HADOOP?

- Hype! Hadoop as the solution to everything
- Operational Complexity
- Evolution (this is positive)
  - Richer programming models
  - Improved faster execution models (memory based)
  - The rise of containers (better support for non Java jobs)
  - The rise of Cloud storage + Datawarehouses

IS HADOOP STILL A THING?

SO HADOOP IS DEAD WHAT NOW?

- HDFS replaced by Cloud Filesystems
- MapReduce replaced by modern programming models e.g. Spark, Beam
- YARN replaced by Kubernetes or by Cloud self-managed systems (Databricks Managed Spark + Google Dataflow)

IS HADOOP STILL A THING?

RISE OF CLOUD SERVICES

To deal with Hadoop operational issues :

- Cheaper and easier storage
- On demand clusters in minutes
  - NY Times transformed 150 years old archive of articles and images into web form on 36h (source)
- Self-managed job execution making operations trivial
- Serverless models (FAAS)

IS HADOOP STILL A THING?

NOT DEAD YET: THE RISE OF THE DATALAKE

- A data lake is a centralized repository that allows you to store all your structured and unstructured data at any scale.

Hadoop’s HDFS and its ecosystem tools fit this definition nicely

CONCLUSION

- L'espace de conception est immense
- Aucun outil ne résout tous les problèmes
- Utiliser le bon outil pour répondre au bon problème :
  - S'il est possible de résoudre un problème simplement, faites-le
  - Si le problème nécessite de multiples sources ou plus de capacité de traitement, utilisez les systèmes distribués
- Chaque projet à son lot de contraintes et compromis
- Enjeux à venir : streaming (depuis 2012)
- Éviter au maximum les systèmes trop complexes
- Les technologies évoluent, mais les principes fondamentaux restent les mêmes

QUESTIONS ?