Fiche de lecture - VIN Charles

J'ai choisi de faire cette fiche en anglais pour éviter les problèmes de traduction

The database community has been impacted by the acceleration of technological breakthroughs in machine learning and data science, the rise of data governance, the growth of managed cloud data systems, the acceleration of the Industrial Internet of Things, and hardware changes. Those advances lead to some new research challenges.

I did not understand all the document but here is my top five, which are linked with our Master and with SAM and that I can relate with.

Data exploration and scale This point discusses the recent advances in machine learning (ML) and how they can be used to reimagine the data platform components. The focus is on auto-tuning. It is described as the way database systems can replace magic numbers and thresholds with ML models, and on exploring new approaches to query optimization or multidimensional index structures. *I like the idea to implement ML in database systems.*

Machine learning workloads This section discusses the importance of ML in modern data management workloads and the need for database engines to support in-database inferencing/training efficiently. To support this, it also highlights the need for database systems to support popular ML programming frameworks such as Pytorch or Tensorflow. This support already exists but only for small ML models, training much bigger AI like language models will require database engine developers to cooperate with data scientists and hardware developers. *I never thought about the way those terabytes of data for language models were managed. I like the way the document said it will need cooperation between researchers and developers of differents fields*

Distributed transactions This section addresses the challenges of processing distributed transactions in cloud data management systems, which are increasingly geo-distributed across multiple geographic regions. There is a debate between two schools of thought, one advocating for reducing consistency and isolation guarantees for high throughput, availability, and low latency, and the other advocating for strong consistency and isolation guarantees. The section emphasizes the importance of better identifying and quantifying application bugs and limitations and building tools to help application developers achieve their goals. *Decentralisation is on-trend and I didn't know that it still was a real problem to have distributed transactions. I think we need to address it quick*

The last two points are around data governance, metadata management and ethical data science. I'll first define those research challenges, them explain the link between them.

- Metadata management involves tracking and managing metadata related to data science experiments and ML models, including automated labelling and annotations of data and data provenance.
- Data governance involves the control of how data is used by applications, with the European Union's General Data Protection Regulation (GDPR) being a prime example.
- Ethical data science involves countering bias and discrimination in the use of data science techniques, with responsible data management emerging as a new research direction in the area of Fairness, Accountability, Transparency, and Ethics (FATE).

I think those challenges are dependent on each other, kind of a domino effect. To improve ethical data science and data governance, we need a database system to be able to keep track of where the data comes from, especially in data lakes that seem to be described as a bit messy. That's where metadata management is involved.