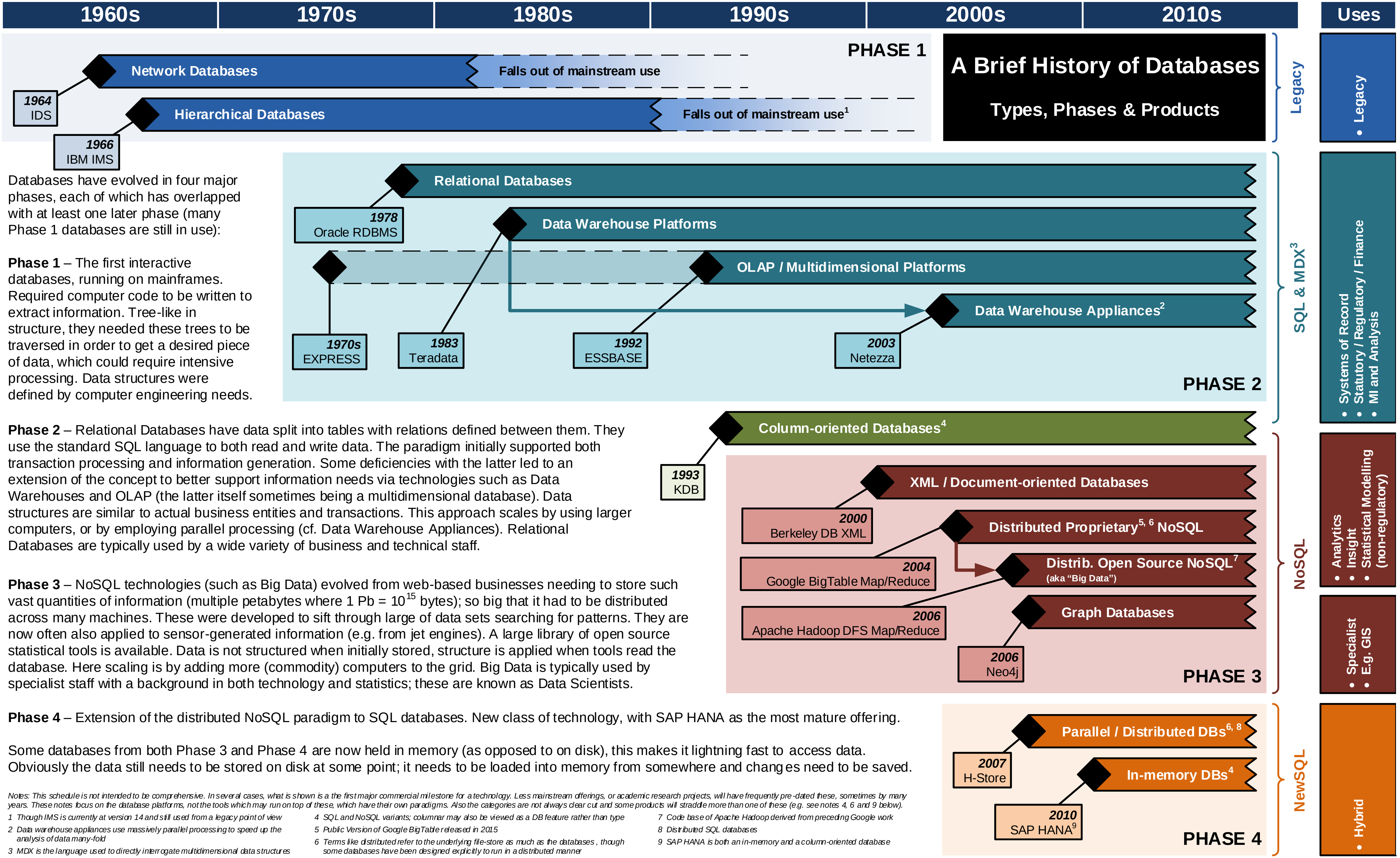
# A Brief History of Databases

## Types, Phases & Products



Databases have evolved in four major phases, each of which has overlapped with at least one later phase (many Phase 1 databases are still in use):

**Phase 1** – The first interactive databases, running on mainframes. Required computer code to be written to extract information. Tree-like in structure, they needed these trees to be traversed in order to get a desired piece of data, which could require intensive processing. Data structures were defined by computer engineering needs.

**Phase 2** – Relational Databases have data split into tables with relations defined between them. They use the standard SQL language to both read and write data. The paradigm initially supported both transaction processing and information generation. Some deficiencies with the latter led to an extension of the concept to better support information needs via technologies such as Data Warehouses and OLAP (the latter itself sometimes being a multidimensional database). Data structures are similar to actual business entities and transactions. This approach scales by using larger computers, or by employing parallel processing (cf. Data Warehouse Appliances). Relational Databases are typically used by a wide variety of business and technical staff.

**Phase 3** – NoSQL technologies (such as Big Data) evolved from web-based businesses needing to store such vast quantities of information (multiple petabytes where 1 Pb = $10^{15}$ bytes); so big that it had to be distributed across many machines. These were developed to sift through large of data sets searching for patterns. They are now often also applied to sensor-generated information (e.g. from jet engines). A large library of open source statistical tools is available. Data is not structured when initially stored, structure is applied when tools read the database. Here scaling is by adding more (commodity) computers to the grid. Big Data is typically used by specialist staff with a background in both technology and statistics; these are known as Data Scientists.

**Phase 4** – Extension of the distributed NoSQL paradigm to SQL databases. New class of technology, with SAP HANA as the most mature offering.

Some databases from both Phase 3 and Phase 4 are now held in memory (as opposed to on disk), this makes it lightning fast to access data. Obviously the data still needs to be stored on disk at some point; it needs to be loaded into memory from somewhere and chang es need to be saved.

*Notes: This schedule is not intended to be comprehensive. In several cases, what is shown is a the first major commercial milestone for a technology. Less mainstream offerings, or academic research projects, will have frequently pre-dated these, sometimes by many years. These notes focus on the database platforms, not the tools which may run on top of these, which have their own paradigms. Also the categories are not always clear cut and some products will straddle more than one of these (e.g. see notes 4, 6 and 9 below).*

*[1] Though IMS is currently at version 14 and still used from a legacy point of view*

*[2] Data warehouse appliances use massively parallel processing to speed up the analysis of data many-fold*

*[3] MDX is the language used to directly interrogate multidimensional data structures*

*[4] SQL and NoSQL variants; columnar may also be viewed as a DB feature rather than type*

*[5] Public Version of Google BigTable released in 2015*

*[6] Terms like distributed refer to the underlying file-store as much as the databases , though some databases have been designed explicitly to run in a distributed manner*

*[7] Code base of Apache Hadoop derived from preceding Google work*

*[8] Distributed SQL databases*

*[9] SAP HANA is both an in-memory and a column-oriented database*

*Acknowledgements:*
*a. Neil Raden (@neilraden) from HiredBrains (hiredbrains.com) – For both reviewing earlier drafts and providing significant input*