Donald D. Chamberlin

This short biography will explore the work and impact of Donald Chamberlin – **co-inventor of SQL**. Today, SQL is the most widely used database language in the world. It was co-developed by Chamberlin and Ray Boyce. It is based around the relational model of data, introduced by Ted Codd in 1970.

We will briefly discuss Chamberlin's early life and education, before examining how he initially came to be involved in the development of SQL. The various twists and turns SQL took along its development trajectory will be explored, and we will examine how Chamberlin (and Boyce's) work fits into the overall development and implementation of the relational database model. Some of Chamberlin's other significant contributions in the area of software engineering will also be highlighted.

Early Life and Education

Donald Chamberlin was born in San Jose, California on December 21 1944. His father was a high-school teacher, who taught English. His mother was a housewife. He attended Campbell High School outside San Jose. According to Chamberlin[2], the launch of the Russian satellite, Sputnik, in 1957, when he was in the eighth grade, made a major impression on him and his generation. It made it clear that technology was important and a national priority. This greatly influenced his early thinking about what he wanted to do.

In 1962, he enrolled in Harvey Mudd College in Claremont, California. Harvey Mudd is a small college that specializes in science and engineering. At that time, they did not yet offer specializations in different types of engineering but rather just a general Bachelor of Engineering degree. This is what Chamberlin graduated with in 1966. From there, he went on to do graduate study at Stanford University on a fellowship from the National Science Foundation (NSF). He considered joining the newly founded Computer Science Department, but he wasn't confident Computer Science would ever be a fully recognized discipline[2]. He instead decided to do his graduate studies in the highly renowned Electrical Engineering Department with Computer Science as a minor. While in graduate school, he interned at several companies, including Hewlett Packard, Lockheed Research Lab in Palo Alto and IBM Research in Yorktown Heights, New York. His doctoral thesis was on a design for a parallel machine.[2] Upon graduating, he joined IBM research in Yorktown Heights full time.

Chamberlin's Work at IBM Research and his Involvement in the System R Project

It was more or less by accident that Chamberlin got involved in database design at IBM. He had originally applied to Yorktown to work on an advanced time-sharing operating system project, called System A. However, System A ended up being consolidated with some product development work at IBM's base in Poughkeepsie. Chamberlin did not want to move to Poughkeepsie and so looked for something else to do in Yorktown.

A charismatic manager, named Leonard Liu, was starting up a database research department at the time. Chamberlin joined. The work involved many parts, including database design work and work on query languages. Chamberlin, who had always been interested in languages, focused on the query language part. So too did Ray Boyce and the pair became close collaborators. They set about studying database query languages, trying to find ways to improve upon the query languages in use at that time.

Navigational database systems were the primary type of database system in use. To answer a query in a navigational database system, you write a command specifying a detailed navigational path to navigate through the pointer paths and find the information you are looking for. This model was problematic as queries not anticipated in advance of the database design could be difficult or impossible to execute.

In June of 1970, Ted Codd, a research fellow at IBM San Jose, published a seminal paper entitled "A Relational Model for Data for Large Shared Data Banks". Codd was a mathematician and knew that all information could be represented by mathematical structures known as relations and queried by expressing questions in a mathematical language known as the relational calculus. One way of conceptualizing a relation is as a table with rows and columns. Codd was of the view that all the information in a database should be represented by such tables. He was also of the firm belief that a query to a database should be expressed as a higher-level question, letting the system figure out how to navigate the database. For what you are trying to do in posing a query is find the answer to a question; the question is of the essence, not the navigation plan. Making the question independent of the plan also allows the computer to change the plan if it becomes more efficient to do so. In addition, it frees the user of the database from having to think about the lower-level procedural details. There was a lot of scepticism at the time that this idea could be efficiently implemented.

After attending a talk given by Codd, Chamberlin and Boyce were converted to the relational point of view. They were taken by the simplicity and elegance of this model. They could see that many queries were easier to express in such a model.

Recognizing that Codd's idea held promise, in 1973 IBM decided to develop an industrial strength prototype as proof of the feasibility of the relational data model. This became known as System R. Chamberlin, Boyce and other IBM researchers, totalling about twelve people, moved to San Jose to become part of the System R team. Chamberlin and Boyce's main job was to design a query language that would constitute the user interface of System R.

They challenged each other in what they called "the query game" to design a language that was flexible enough to express many types of queries: One person would dream up a question and the other would try to find a way to express this question in a computer language.

They studied the relational languages put forward by Codd, which they deemed to have a number of shortcomings. For one thing, they were couched in mathematical notation, such as universal and existential quantifiers. This was off-putting to non-mathematician practitioners and could not be typed on a keyboard. Another issue was that it was only possible to express queries; Codd's languages didn't have any capabilities for updating data.

In real world applications, such as a banking or airline reservation, continuous updates must be made to the data. So Chamberlin and Boyce decided this needed to be an integrated part of the language.

In fact, they decided that all tasks involved in database management (including inserting, deleting and updating data, controlling access authorization and defining constraints to maintain database integrity) should be able to be accomplished by a single language in a uniform syntactic framework.[1]

They wanted their language to be based on English keywords and, as Codd had proposed, to be declarative rather than procedural. They wanted to capture the power and simplicity of the relational model, whilst avoiding the mathematical terminology of Codd's original proposal.

Phyllis Reisner was a cognitive psychologist at IBM. She carried out a series of human factors experiments on college students to test the learnability of some of Chamberlin and Boyce's early proposals for such a query language.

In May 1974, Chamberlin and Boyce published a 16-page paper containing an initial proposal for a query language, based around English keywords, called **SEQUEL** (an acronym for "**S**tructured **E**nglish **Q**uery **L**anguage").

Sadly, Ray Boyce passed away suddenly as a result of a brain aneurysm shortly thereafter.

After this initial proposal, the SEQUEL language went through a validation and refinement phase that lasted from 1974 to 1979. During this phase, the language was implemented and the implementation process fed back into the evolving design of SEQUEL. A prototype was installed for free at three customer locations, who used the language on an experimental basis and provided feedback on problems and possible improvements.[2]

The name of the language had to be changed from SEQUEL to SQL to avoid a trademark infringement. [1]

The first commercial version of SQL was released in 1979. Today, many implementations of SQL are available across all major platforms.

The impact of SQL

The development of SQL by Chamberlin and Boyce greatly helped popularize Codd's Relational Model, making it more easily understandable to the non-mathematician. The development of a language based on English keywords, which could be intuitively understood, was a major breakthrough. It didn't actually make the underlying ideas any more simple; it just made them look simpler.[2]

On its own, however, SQL would not have been sufficient to fully realise Codd's ideas. It is but a piece of a larger puzzle. A number of other components were also necessary to successfully implement the Relational Model. Another crucial problem to be solved was to find a way to automatically generate an efficient navigational plan from a given higher-level query in a

language like SQL. Pat Selinger provided a solution to this difficult problem by designing a cost-based query optimizer for System R. [2] Her solution gave people some confidence that relational database systems would be able to perform at a level on par with their navigational counterparts. Raymond Lorie, another member of the System R team, developed a query compiler, which, for frequently posed queries, optimized the query just once and saved the navigational plan[2]. Jim Gray, on the other hand, studied the theory of transactions in a rigorous way, allowing database transactions to be implemented in a safe and consistent manner.[2] Codd, Chamberlin, Boyce, Selinger, Lorie, Gray, and many others we don't have time to mention here, together laid the foundations for the success of relational databases. This illustrates how the impact of Chamberlin and Boyce's work depended on the work of many others, and how the work of others in turn depended on Chamberlin and Boyce's work. This is typical of Software Engineering. It is a team endeavour.

The relational model subsequently took over the world of commercial data processing and is still an extremely widely used data model today.

Chamberlin's Other Contributions to Software Engineering

Later in his career, Chamberlin studied the problem of how to query data in the XML format. He represented IBM on the W3C working group for XML Query from the time of that group's formation in 1999. He put forward a proposal for an XML query language, called Quilt. Many of his ideas made it into the XQuery language specification.

Awards and Accolades

In 1988, Chamberlin was (jointly) awarded the *ACM Software Systems Award* for his work on System R.

In 2009, he was made a Fellow of the *Computer History Museum* "for his fundamental work on structured query language (SQL) and database architectures."

Bibliography

1. Biancuzzi, F., & Warden, S. (2009). *Masterminds of programming* (1st ed.). Sebastopol, CA: O'Reilly.

2. Oral History Interview with Donald D. Chamberlin. Conducted by the Charles Babbage Institute. San Jose, California (2001). Available at http://hdl.handle.net/11299/107215