27 Dimensionalität des optimalen Datenraumes

27.1 Definitionen:

Datenraum: In der praktischen Informatik ist der Datenraum eine der wichtigsten Größen. Er beschreibt jenen Raum, der von den Einzeldaten gefüllt wird und in seiner Gesamtheit den Inhalt eines Speichers beschreibt. Aufgrund der Abstraktheit der Weltabbildung im Rechner kann der Datenraum in seiner Dimensionalität vom Programmierer einer Datenbank in weiten Bereichen frei bestimmt werden. Von diesen Bestimmungen hängen so wesentliche wirtschaftliche Faktoren wie Rechenzeiten, Programmierkosten, Einsatzfähigkeit und Systemkosten der EDV ab.

Dimensionalität: Datenräume sind physisch immer eindimensional wegen der grundsätzlich seriellen Arbeitsweise der derzeit verwendeten Prozessoren. Aufgrund der strukturellen Eigenheiten der Datenverarbeitung (näheres siehe Kotauczek, Die Weltbildmaschine) kann mit geeigneter Programmierung jede Dimensionalität im Rechner emuliert werden. Dabei sind auch gebrochene (fraktale) Dimensionen (i.S Haussdorf/Mandelbrot) möglich und werden auch täglich verwendet.

Optimaler Datenraum: Je höher die Dimensionalität desto „löchriger“ wird die Datenbesetzung in einem gegebenen Array, das hat bisweilen dramatische Auswirkungen auf den Speicherbedarf und die Zugriffszeiten. Zum Zeitpunkt dieses Berichtes leidet gerade das Netzwerk der BEKO-Gruppe stark unter diesem Syndrom, was offenbar eine Folge historischer Fehlentwicklungen und versäumter Optimierungsbemühungen sein dürfte. Die Suche nach dem theoretisch optimalen Datenraum ist daher eine ständige Aufgabe für die das IHI gerüstet sein möchte.

Lakunarität: die von dem Mathematiker Mandelbrot gefundene Löchrigkeit bei gebrochenen Dimensionen wie der Cantor-Menge, dem Sierpinsky-Dreieck oder dem Fatou-Staub. Die Künstler und Wissenschaftler, die sich mit Fraktalität beschäftigen sind mit diesem Phänomen bestens vertraut. Das berühmte „Apfelmännchen“ der Mandelbrot-Menge weist auch Lakunaritäten auf.

27.2 Grundlagen

Die Datenverarbeitung hat die Aufgabe, Sachverhalte aus der Realwelt so abzubilden, dass der Nutzer konkrete praktische Erkenntnisse für seine Entscheidungen daraus ziehen kann. Weicht die im Datenwürfel abgebildete Teilrealität zu weit von der realen Wirklichkeit ab, dann kommt es zwangsläufig zu Fehlentscheidungen. Das gilt unabhängig von der fachlichen Ausrichtung des jeweiligen Datenbestandes.

Nun ist ja bekannt, dass die meisten Anwendung heute auf bestehenden Datenbanksystemen aufgesetzt werden Man könnte daher annehmen, dass die Frage des optimalen Datenkubus längst gelöst sei. Das stimmt aber leider nicht. Gerade die Flexibilität der angebotenen Datenbanksysteme und der zugehörigen OLAP-Applikationen bieten so eine reichhaltige Wahlmöglichkeit der verwendeten Dimensionen, dass dieser Umstand zu einer zunehmenden Verschwendung von Datenraum führt, die fast schon an Ressourcenmissbrauch grenzt.

27.3 Der IHI-Forschungsansatz

Um festzustellen, wie weit sich ein realisiertes Datenbankprojekt vom Optimum entfernt hat, braucht es ein theoretisches Fundament auf das man die realisierte Lösung vergleichsmäßig beziehen kann. Anders kann man nicht feststellen, wie sub/optimal ein bestehendes System arbeitet.
Dabei ist aber immer zu beachten, dass jedes praktisch implementierte System immer suboptimal ist, weil technische, finanzielle oder personelle Limits bei der Planung, Programmierung und Praxiseinführung in jedem Teilsystem/Modul zur ungewollten Abweichung vom theoretisch möglichen Optimum führen. Die Summe aller Abweichungen führen letztendlich zur Gesamtabweichung und definieren die Kosten/Nutzen-Relation des Systems im praktischen Alltag.
Das IHI versucht daher seit Jahren, in der Literatur und im direkten Kontakt mit Fachleuten aus den verschiedensten Disziplinen herauszufinden, ob es ein Optimum in der Dimensionalität der Datenräume überhaupt gibt und wie dieses aussehen könnte.

27.4 Die Pi-Dimensionalität

Die Fachwelt ist sich weitestgehend darüber einig, dass die Welt mindestens dreidimensional organisiert ist. Die gesamte Zeit in der die Newtonsche Physikauffassung die Wissenschaft prägte war die Dreidimensionalität ein Dogma. Erst durch Einstein kam die Zeit als vierte Dimension so richtig ins Spiel und ist seither Faktum des Schulwissens. Jeder kennt das Wort vom „Raumzeitkontinuum“ der Allgemeinen Relativitätstheorie.
Mit dem Aufkommen der Quantentheorie kam der Grundgedanke der „Sprunghaftigkeit“ der Natur ins Bewusstsein der Wissenschaft und musste nach anfänglichen Versuchen, sie als reines Gedankenkonstrukt abzulehnen, unter dem Druck der experimentellen Befunde als gleichermaßen richtig anerkannt werden.
Seither gibt es immer wieder Versuche einer Vereinigung dieser beiden physikalischen Weltbeschreibungen. Die derzeit am meiste diskutierte ist die Stringtheorie mit ihren bis zu 26 Dimensionen (bosonische Stringtheorie).
Was heißt das für die praktische Arbeit des IHI? Alle Befunde, die dem IHI vorliegen, deuten in die Richtung, dass es einen Trade-Off zwischen Dimensionalität und Lakunarität gibt. In anderen Worten, je mehr Dimensionen ein endlicher Datenraum aufweist desto grösser müssen die „Löcher“ zwischen den Datenclustern werden. Das ist auch die tägliche Beobachtung in der IT-Praxis, auch und nicht zuletzt im BEKO-Datennetzwerk. Wenn die Lakunarität aber zu groß wird, muss der Computer (Server) fast nur mehr unbesetzte Speicherstellen in Form von Leerschleifen abgrasen, um irgendwann endlich ein brauchbares Datenfragment zu finden. Das kostet Zeit und Geld. Eine durchschnittlich 10%ige Lakunarität pro Dimension führt bei 26 Dimensionen (wie bei der oben genannten Bosonen-Stringtheorie) bereits zu einer Dichte im Datenbestand von 0,9**26 = 0,06461. Nur mehr 6,5 % des Speicherplatzes sind noch mit relevanten Daten befüllt, der 93,5 % Rest ist leerer Speicherraum.
Um das zu vermeiden gibt es seit jeher Bemühungen die Lakunarität zu vermindern. Ein bewährter Trick in der IT beruht darauf, alle Leerstellen wegzulassen aber sie sich gleichzeitig irgendwie zu merken um sie später wieder am richtigen Platz einfügen zu können. Das knifflige Problem dabei ist das sichere Auseinanderhalten zwischen leerer Speicherzelle und einer 0-Speicherzelle. Diese Vorgangsweise bringt einen hohen Verdichtungseffekt am Datenträger. Aber wo gibt es eine Grenze?
Verschiedene Überlegungen haben das IHI dazu gebracht, die Vermutung zu postulieren die Summe aller Dimensionen in einem maximal verdichteten Datenraum dürfe Pi nicht überschreiten.

$\Sigma D = \pi$
das ergibt abgeleitet:
$\varnothing d = \pi / D$

In Worten: die durchschnittliche Dimension d ist Pi dividiert durch die Anzahl der Dimensionen des betrachteten Datenraums. Da Pi nicht ganzzahlig teilbar ist muss es mindestens eine gebrochene Dimension geben.
Bisher war es dem IHI nicht möglich jemanden zu finden, der eine ähnliche oder gar identische Auffassung vertritt. Immerhin setzt dieses Postulat voraus, dass man die Existenz nicht ganzzahliger Dimensionen auch in der Realität anerkennt. Die immer grösser werdende Community der Anhänger der Fraktalität in Kunst und Wissenschaft tut das.
Die Kunst ist wieder einmal etwas vorne weg, weil sie das Privileg genießt, spekulative Annahmen ohne Bremsung durch die kanonische Lehre ausprobieren zu dürfen. So ist das auch im Alltag des IHI. Als Kunstprojekt getarnt konnte CALSI jahrelang innerhalb der BEKO als Projekt überleben unabhängig von den gerade amtierenden Managements bzw. wechselnden Eigentümerstrukturen. Genau so oder ähnlich läuft das auch in der gesamten österreichischen IT-Industrie.

27.5 Weiterführende Überlegungen

Warum gerade Pi als Summe der Dimensionen? Nun, ein Datenraum, der die Realwelt möglichst 100% abbilden soll, muss die gleiche Dimensionalität aufweisen wie die Welt die abgebildet werden soll. Also salopp gesprochen, die ganze Welt wie sie liegt und steht. Nur dann kann die EDV für sich beanspruchen, potentiell ein exaktes Abbild der Welt zu liefern. Und genau das wird von den Usern gefordert.
Vor kurzem ist es dem IHI-Knowledge-Netzwerk gelungen, eine profunde Deduktion von Galvin Roy Fox in die Hand zu bekommen, die den Titel trägt (übersetzt):
„Wie ich (CRF) deduzieren konnte, dass das Universum Pi-Dimensional ist“
Diese Arbeit kann über das IHI im Original angefordert werden, sodass hier eine verkürzte Darstellung (Management-Summary) der Gedanken des Autors ausreicht.
Fox behauptet es sei eine Sache der Definition. Er argumentiert, durch Beobachtung des Gesetzes der Kausalität könne man schließen, das Universum sei der Effekt von „Was-auch-immer“ es verursacht hat zu passieren und zu existieren.
Das was existiert ist definiert durch seine Existenz, unabhängig davon was es sein mag. Das Universum sei die Existenz von Irgendetwas und daher die Definition dieses Irgendetwas, so die Argumentation von Fox.
Nun könnte man einwenden, das sei eine tautologische Betrachtungsweise. Der Einwand wird vermutlich auch kommen. Aber es ist eine typisch IT-übliche Aussage. Die ganze IT lebt von dieser Wittgensteinschen Denkfigur (Die Welt ist alles was der Fall ist). Der Pancomputationalismus eines Seth Lloyd oder Stephen Wolfram belegt das ausführlich.
Zu fragen, was die Zeit selbst verursacht hat zu beginnen, führt Fox weiter aus, trägt die Annahme in sich, es gäbe eine lineare Basis, nach der der Effekt immer nach der Ursache kommt. (Das IHI hat im 35. IHI-Bericht auf diese versteckte Prämisse im Wissenschaftsbetrieb hingewiesen und die Kritik von Wolfram daran kurz beschrieben). Fox verweist auf die Tatsache, dass die Annahme eines linearen Kausalnexus zwangsläufig zu einem endlosen Regress führt, weil man ja bei jedem hypothetischen Zeitbeginn wiederum die Kausalitätsfrage stellen kann: was war die Ursache der „Zeit“ vor der Zeit usw.
Fox verweist darauf, dass das Problem damit geschaffen wird, wenn man die Zeit als Effekt sieht. Der einzige Weg, dieser Logikschleife zu entkommen, sei laut Fox die Annahme, die Zeit sei kein Effekt, sondern die Abwesenheit eines Effektes aufgrund der Abwesenheit einer Ursache. Am Beginn der Zeit war gar nichts. Wie in der IT: bevor Daten in den Computer geladen werden, ist da gar nichts. Keine Ursache bedeutet kein Effekt.
Fox stellt die rhetorische Frage: keine Ursache daher kein Effekt, also existiert nichts und frägt: ist das möglich?
Und er schließt daraus: es ist nicht möglich da die Welt ja existiert. Daher bleibt die Möglichkeit (der Existenz). Auf den Computer umgelegt ist das die Frage des Unterschiedes zwischen 0 und Nichts ist gleich 0-Datenzelle versus leere Datenzelle versus nicht vorhandene Datenzelle. Mit der implizit

eingeführten Menge aller Möglichkeiten nähert sich Fox der von Zadeh eingeführten Theorie der unscharfen Mengen, besser bekannt unter dem Schlagwort „Fuzzy Logic“ an, wo auch zwischen absoluter Möglichkeit (Gewissheit) mit dem Wert 1 und absoluter Unmöglichkeit (Wert 0) operiert wird.
Fox nimmt nun die Überlegung, dass „die Möglichkeit des Nichts“ alles ausschließt was existiert als das Zentrum der „Absenz von Allem“. Er sagt, was immer von diesem Zentrum der Absenz in Richtung Existenz ausgeht, egal in welche dimensionale Richtung es gehen mag sei da und endlich. Nach guter mathematischer und IT-Sitte misst er dem „Zentrum der Absenz“ den Wert 0 zu und allen anderen Abständen des Seins den Wert 1. Und da die Richtung beliebig und der Wert immer maximal 1 ist, nennt er das was da als abstrakter Graph herauskommt „rund“. Wie man einen Kreis rund nennt, wenn der Abstand um das Zentrum immer 1 (also endlich) ist.
Nun, so die zentrale Schlussfolgerung von Fox, wenn etwas rund, endlich und von der Quantität 1 ist, dann ist der einzige absolut gesicherte Aspekt, dass dieses Etwas nicht absolut sicher in seiner Dimensionalität sein kann. Es kann zwar nicht weniger als eine Dimension haben, aber weil es anderseits keine absolute Dimension haben kann muss es mehr als eine Dimension haben und die totale Zahl der Dimensionen muss endlich sein.
Daher meint auch Fox so wie das IHI, weil das, was ist, von der Dimensionalität grösser als 1 sein muss, rund ist und endlich ist, wäre Pi der beste Aspekt seiner Existenz. Das Seiende (die Welt, das Universum, die Grundgesamtheit) sei Pi-Dimensional.
Was heißt das nun für die Praxis? Das Pi-dimensionale Universum wäre ein ideal gepacktes Real-„Kontinuum“. Es entspricht einem vollgepackten lückenfreien Speicherraum im Computer. Mit diesem kann man sehr gut ganzzahlige Data-Cubes abbilden. Das setzt aber voraus, dass die Gesamtheit der Dimensionalität auch ein ganzzahliges Vielfaches von 1 ist. Das ist bei einem dreidimensionalen Raum, der lückenfrei wie der Newton-Raum ist, der Fall. Bei einem gekrümmten Einstein-Raum geht das nicht mehr und noch weniger bei einem Stringtheorie-Raum. Solche Raumkonzepte sind nur mehr lokal als homogen und isotrop konstruierbar. Die Wissenschaft weiß das natürlich und spricht daher immer von einer (Riemann-geometrischen) Mannigfaltigkeit.
Aber die Bedeutung der Pi-Dimensionalität geht viel weiter. So ist in der Bildverarbeitung der Einfluss der Abtastraten auf die Bildqualität genau so bekannt wie im Finanzwesen der Einfluss der Erfassungsperioden auf die Periodenergebnisse. Jedem CFO ist das Spiel mit der Stichtagserfassung für die Bilanzgestaltung bekannt oder die Dehnung der Bewertungs- „Räume“ von Assets. Ein guter Teil der derzeitigen Finanzkrise geht auf unstatthafte Manipulationen von Datenräumen und der damit verbundenen Fehlallokationen von Einzeldaten in bedeutungstragenden Ergebnisräumen einher. Organisationen wie IFRS oder Rating-Agenturen mühen sich mit dieser Problematik ständig ab und beeinflussen das Schicksal von Firmen und ganzer Völker.
Auch im politischen Bereich ist längst bekannt, dass man durch selektives Feinstellen von Beobachtungsrastern Datenbedeutungen manipulieren kann ohne die Originaldaten verfälschen zu müssen.
Würden die Erkenntnisse der Dimensionalitätsforschung besser bekannt sein und in einer breiteren Elite diskutiert werden, wüsste jeder, dass eine datenmäßige Abbildung der Welt zwangsläufig immer Lücken aufweisen muss und es nichts nützen kann, immer noch mehr in Datenerfassung und Datenhaltung zu investieren ohne gleichzeitig die Lakunarität gebrochener Dimensionen zu berücksichtigen.
Je genauer man eine Dimension datenerfassend verfolgt desto größere Lücken tun sich zwangsläufig bei den anderen Dimensionen auf. Je höher die angewendete Dimensionalität des Datenraumes desto kleiner wird der Wert der Einzeldimension bzw. desto grösser der Lückenanteil. Allerdings könne einzelne Dimensionen gegen 1 gehen also annähernd lückenfrei auf Kosten der anderen Dimensionen werden.

40. IHI Bericht, 27.01.2012