Das Data Warehouse der BA

Information im Rahmen des Workshop zur Evaluation der Arbeitsförderung

Hans-Peter Schwarzfärber (BA, IT 43)

Das Data Warehouse der BA

Information im Rahmen des Workshop zur Evaluation der Arbeitsförderung

INHALTSVERZEICHNIS

1 Was ist das Data Warehouse der BA?

Die Bundesanstalt verfügt über eine Vielzahl von Datenbeständen, die durch operative IT-Verfahren (z.B. coArb, coLei, coSach) bearbeitet werden. Da Daten zum gleichen Sachverhalt in den operativen IT-Verfahren bisher sowohl zeitlich, inhaltlich als auch örtlich unterschiedlich erfasst werden, ist es oft nicht möglich, Daten zur Unterstützung geschäftspolitischer Entscheidungen sinnvoll zusammenzuführen. Zur Optimierung von Geschäftsprozessen, zur besseren Berücksichtigung regionaler Gegebenheiten bzw. ganz allgemein zur Verbesserung der Leistungen der Bundesanstalt ist aber genau dieses erstrebenswert.

An dieser Stelle setzt nun das Data Warehouse an. Vielfach wird der amerikanische Begriff Data Warehouse mit „Daten Warenhaus“ übersetzt und vermittelt damit das Bild eines Selbstbedienungsladens, in dem Informationen übersichtlich und ordentlich aufgereiht für den Endbenutzer zum Abholen bereitliegen. Ein Blick in das Wörterbuch zeigt, dass der Begriff Warehouse für Lager oder Speicher steht. Diese Übersetzung kommt der Wirklichkeit viel näher. Genau genommen handelt es sich bei einem Data Warehouse um eine zentrale Datenbasis, in die Daten aus unterschiedlichen operativen internen Systemen aber auch von externen Stellen, z.B. Sozialversicherungsträgern oder statistischem Bundesamt, geprüft, aufeinander abgestimmt und zusammengefasst werden. Die Daten werden in dieser einheitlichen Datenbasis (dem Datenlager) mit dem Ziel abgelegt, den Entscheidungsträgern auf Anforderung schnell Daten zur Überprüfung von getroffenen oder als Hinweis auf anstehende Entscheidungen zu liefern.

Im Zusammenhang mit Data Warehouse taucht auch immer wieder der Begriff Data Mart auf. Bei einem Data Mart handelt es sich um eine Teilmenge eines Data Warehouse. Durch die zeitliche Fortschreibung von Daten (Historisierung) und optimierende Zusammenfassungen (Aggregationen) erreicht ein Data Warehouse schnell Speichermengen in der Größenordnung von einigen hundert Gigabyte (1 Gigabyte = 1 Milliarde Byte) bis zu mehreren Terabyte (1 Terabyte = 1000 Gigabyte). Ein Anwender bzw. eine Anwendergruppe benötigt in den meisten Fällen aber nur einen fachlich abgegrenzten Teil der in einem Data Warehouse abgelegten Daten. Diese nach fachlichen Gesichtpunkten organisierten Daten werden deshalb in einem Data Mart den jeweiligen Benutzern separat zur Verfügung gestellt. Im Laufe der Zeit entstehen deshalb um ein Data Warehouse herum immer mehr Data Marts. Eine große Herausforderung für die Verwaltung des Gesamtsystems Data Warehouse besteht nun gerade darin, den Gleichlauf des Data Warehouse mit den Data Marts sicherzustellen.

Fest mit dem Begriff Data Warehouse ist auch die Bezeichnung OLAP verbunden. OLAP steht dabei für Online Analytical Processing und beschreibt eine Software-Technologie, die schnelle und vielfältige Dialogzugriffe auf die in einem Data Warehouse oder auch einem Data Mart abgelegten Informationen gestattet. Im Grunde genommen handelt es sich dabei um hochentwickelte Abfrageprogramme, die auf die spezielle Struktur bzw. die Anforderungen eines Data Warehouse ausgerichtet sind und dem Endanwender den Zugriff auf vorgefertigte Berichte oder die Zusammenstellung komplexer Abfragen der Datenbestände ermöglichen.

2 Ziele des Data Warehouse der BA

Die BA hat sich mit dem Data Warehouse zum Ziel gesetzt, die Qualität und die Auswertungsmöglichkeiten der geschäftspolitisch relevanten (dispositiven) Daten nachhaltig zu verbessern. Das bedeutet insbesondere:

- **Basiskonsolidierung der Daten**
  Die derzeit noch isolierten Datenbestände sollen zu einem (logischen) Datenbestand zusammengeführt werden. Dies bedeutet insbesondere, dass alle auswertungsrelevanten Datensätze des Data Warehouse den übergeordneten Datenobjekten „Person“ und „Betrieb“ zugeordnet werden. Diese konsolidierte Datenbasis ist die Grundlage, um fachübergreifende Erkenntnisse und Zusammenhänge gewinnen zu können.

- **Datenqualität**

  Alle internen und externen Daten, die in das Data Warehouse übernommen werden sollen, durchlaufen fachlich abgestimmte Prüfungen, die Widersprüche erkennen und ggf. beseitigen können.

- **Kostenreduktion**

  Für die Erzeugung von Auswertungen (z.B. statistische Tabellen) werden Programme eingesetzt, die den Endanwendern eine grafische Oberfläche bieten, um die auszuwertenden Daten zusammenzustellen und das Layout der Ergebnistabellen zu gestalten. Auf diese Weise können Berichte schneller und somit kostengünstiger erzeugt werden.

- **Zugriffsoptimierung**

  Vorgefertigte Berichte, die regelmäßig mit gleichem Inhalt und Layout benötigt werden, können durch Web-Browser (Intra-/Internet) abgerufen werden. Es besteht die Möglichkeit, die Ergebniszahlen direkt in andere Bürokommunikationsprodukte (z.B. Excel) zu übernehmen.

- **Metadatenbasis**

  Die Beschreibung aller Komponenten und Ereignisse, die im Umfeld des Data Warehouse eine Rolle spielen, wird in einer Metadatenbasis organisiert. Es werden technische und fachliche Metadaten unterschieden. Die technischen Metadaten beschreiben unter anderem die Daten, die Laderegeln und die Datenstände des Data Warehouse. Die fachlichen Metadaten hingegen beziehen sich insbesondere auf Begriffsdefinitionen, auf die Dokumentation vorgefertigter Ergebnisse und auf Erläuterungen von Besonderheiten von Berichtsergebnisseen.

- **Reverse-Engineering**

  Die Funktionalität der bestehenden Verfahren Beschäftigtenstatistik (BS2000), STADA und coStat soll durch das Data Warehouse abgelöst werden.

3 Komponenten des Data Warehouse der BA

3.1 Allgemeine Architektur

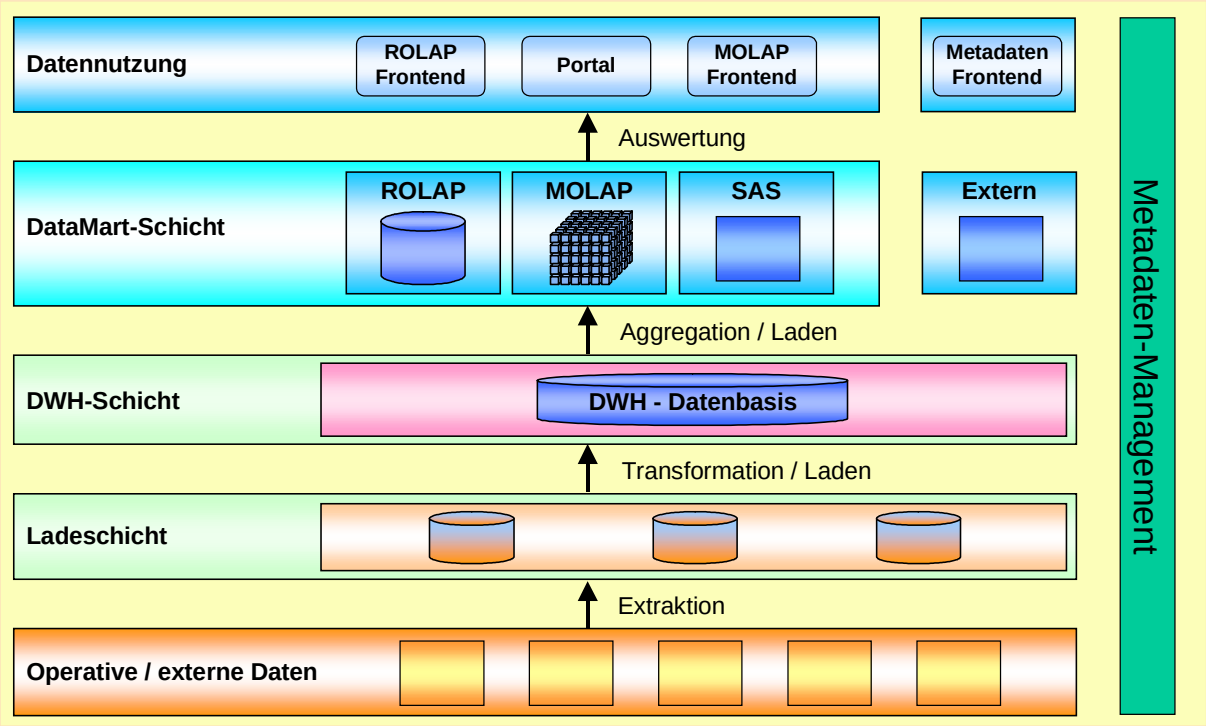


Bild1: Allgemeine Architektur des Data Warehouse der BA

- **Operative / externe Daten**

  Die Daten, die in das Data Warehouse der BA übernommen werden sollen, kommen aus den operativen IT-Verfahren der BA (z.B. Arbeitsvermittlung, Ausbildungsstellenvermittlung, Leistung, Orts- und Betriebsdatenverwaltung) oder von externen Stellen (z.B. Verband der Rentenversicherungsträger, Statistisches Bundesamt). Die Daten werden in der Regel per Datenübertragung (in Ausnahmefällen per Datenträger) angeliefert und als ASCII-Dateien im Filesystem des DWH-Servers zwischengespeichert.

- **Ladeschicht**

  Die im Filesystem bereitgestellten operativen / externen Daten werden syntaktisch geprüft (z.B. Prüfung auf gültige Datumswerte, Versicherungsnummer, Ortsangaben) und in das Datenbanksystem des Data Warehouse geladen. Fehlerhafte Datenfelder werden gekennzeichnet. Dieser Schritt wird als **Extraktion** bezeichnet.

- **DWH-Schicht**

  Die Daten der Ladeschicht werden konsolidiert, indem sie in die relationale Datenstruktur des Data Warehouse integriert werden. Integration bedeutet, dass die Daten

  - den übergeordneten Datenobjekten „Person“ und „Betrieb“ zugeordnet
  - in der korrekten historischen Datenfolge eingereiht
  - durch entsprechende Schlüsselwerte (Fremdschlüssel) der Dimensionen ersetzt

  werden. Darüber hinaus werden komplexe Prüfungen (z.B. Ermittlung von Gültigkeitszeiträumen) durchlaufen. Fehlerhafte Datenfelder werden gekennzeichnet. Dieser Schritt wird als **Transformation / Laden** bezeichnet.

  In der DWH-Schicht liegen alle Einzeldaten des Data Warehouse in konsolidierter Form vor. Die DWH-Schicht ist sozusagen das Zentrum des Data Warehouse (DWH-Datenbasis), aus dem sich die fachlich geforderten Auswertungen gewinnen lassen.

- **DataMart-Schicht**

  Da nur wenige Endanwender die Sicht auf alle Datenobjekte haben müssen, werden die Daten der DWH-Schicht nach fachlichen und technischen Gesichtspunkten aggregiert und in die DataMart-Schicht übernommen. Dieser Schritt wird als **Aggregation / Laden** bezeichnet.

  Unter fachlicher Aggregation versteht man, dass nur die Datensätze selektiert werden, die für einen bestimmten Stichtag oder Zeitraum relevant sind. Unter technischer Aggregation versteht man, dass häufig benötigte Auswertungen bereits vorberechnet werden (z.B. Zusammenfassung von Daten nach Arbeitsämtern). Die technische Aggregation dient zur Verbesserung der Performance.

- **Datennutzung**

  Die Datennutzung ermöglicht über ein Portal den gesicherten Zugriff auf das Data Warehouse und dessen Metadaten. Darüber hinaus wird eine grafische Oberfläche zur Verfügung gestellt, mit der fertige Berichte abgerufen, Daten ausgewertet und Auswertungsergebnisse aufbereitet werden können. Dieser Schritt wird als **Auswertung** bezeichnet.

3.2 Hardware und betriebssystemnahe Software

3.2.1 Zentrale UNIX-Systeme (ZeUS)

Auf den zentralen UNIX-Systemen liegt die Ladeschicht und die DWH-Schicht (Datenbasis) des Data Warehouse in Form eines Datenbankmanagementsystems der Fa. Informix Extended Parallel Server (XPS). Bei relationalen Auswertungen (ROLAP) wird das Datenbankmanagementsystem auch für die DataMart-Schicht verwendet.

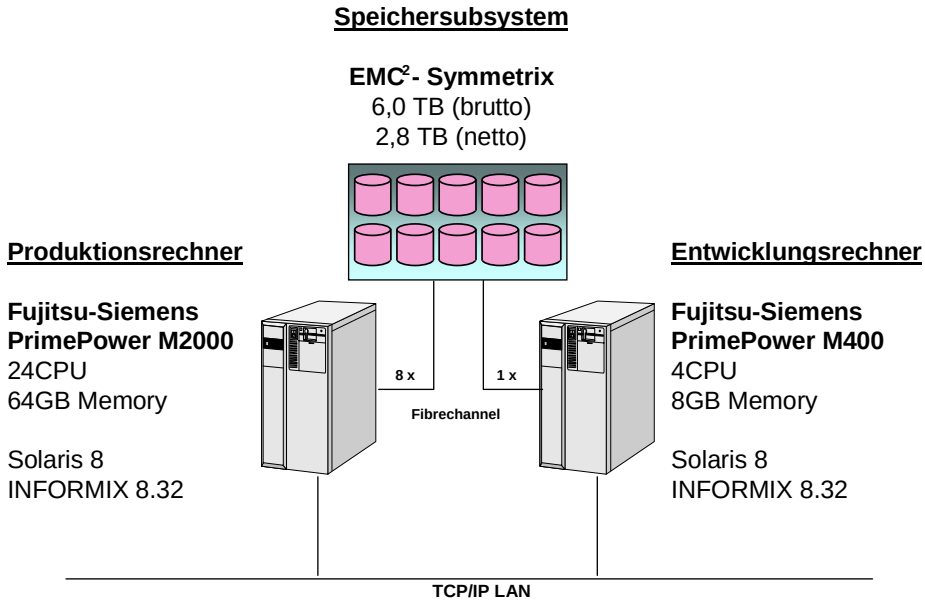
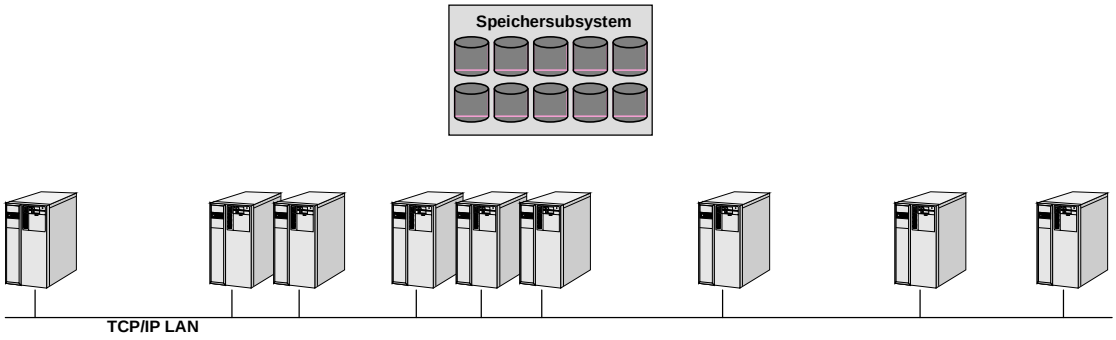


Bild 2: Zentrale UNIX-Systeme des Data Warehouse der BA

3.2.2 Zentrale Windows-Systeme

Auf den WindowsNT- und Windows2000-Systemen liegt die DataMart-Schicht des Data Warehouse in der Form von Datenwürfeln der multidimensionalen Datenbasis Analysis Server 2000 der Fa. Microsoft. Darüber hinaus sind dort zugriffsunterstützende Prozesse/Produkte (Relationaler Server, Metadaten-Server, Web-Server) und die Metadaten angesiedelt.



Relationaler Server	Multidimensionale (Lade-)Server	Multidimensionale (Query-)Server	Metadaten Server	Relationaler Intranet-Server	Multidimensionaler Intranet-Server
Windows2000-Server 4CPU 4GB Memory	**2 Windows2000-Server** 8CPU 8GB Memory	**3 Windows2000-Server** 8CPU 8GB Memory	**Windows2000-Server** 8CPU 8GB Memory	**NT-Server** 4CPU 4GB Memory	**NT-Server** 4CPU 4GB Memory
Windows 2000 SQL-Server 2000	Windows 2000 SQL-Server 2000	Windows 2000 SQL-Server 2000	Windows 2000 SQL-Server 2000	Windows 2000	Windows 2000

Bild 3: Zentrale Windows Systeme des Data Warehouse der BA

3.2.3 Client-Systeme

Die Client-Systeme sind in der Regel Standard-PC's, die in einer Dienststelle der BA (AÄ, LAÄ, HSt) angeschlossen sind. Der Zugriff auf die Daten und Ergebnisse des Data Warehouse erfolgt entweder direkt aus dem OLAP-Tool (dazu muss das OLAP-Tool auf dem Client-System installiert sein) oder über das Intranet (dann genügt der Web-Browser Internet Explorer). Beim Web-Zugriff ist für den Endanwender in Teilbereichen eine eingeschränkte Funktionalität verfügbar.

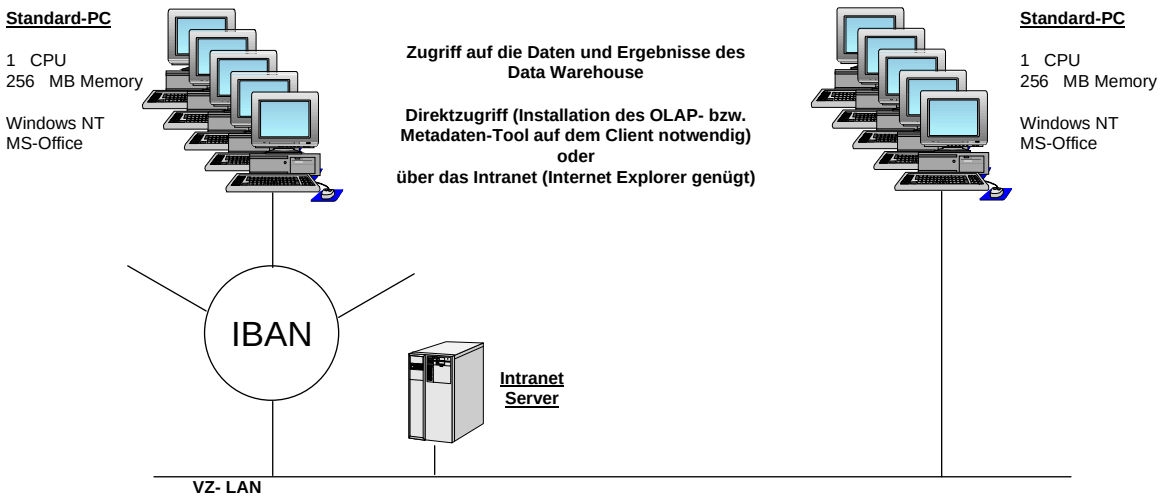


Bild 4: Client Systeme des Data Warehouse der BA

3.3 Software

Für den Betrieb des Data Warehouse der BA werden folgende Standard-Software-Produkte benötigt:

- **CASE-Tool**

  Das CASE-Tool dient zur strukturierten Beschreibung des Datenmodells der DWH-Datenbasis und der fachlichen Anforderungen, nach welchen Regeln die Daten in das Data Warehouse übernommen werden sollen.

  Bei der BA wird das CASE-Tool Innovator der Fa. MID eingesetzt.

- **Datenbanksystem**

  Das Datenbanksystem dient zur Datenkonsolidierung, zu Datenqualitätsverbesserungen und zur Speicherung der Einzeldatensätze des Data Warehouse.

  Bei der BA wird das relationale Datenbanksystem Extended Parallel Server der Fa. Informix eingesetzt.

- **Multidimensionale Datenhaltung**

  Die multidimensionale Datenhaltung ist der Datenspeicher der Data Marts. In den Data Marts werden die Ergebnisse der nach fachlichen Gesichtspunkten organisierten Einzeldatensätze des Data Warehouse zusammengefasst. Auf diese Weise kann insbesondere die Datenbasis für Standardberichte bereitgestellt werden, die für die Endanwender leicht abrufbar sind.

  Bei der BA wird die multidimensionale Datenhaltung Analysis Server 2000 der Fa. Microsoft eingesetzt.

- **ETL-Tool**

  Das ETL-Tool dient zur Unterstützung der Extraktion bzw. Transformation der Daten aus den operativen IT-Verfahren der BA oder externen Datenquellen und zum Laden dieser Daten in das Datenbanksystem und in die multidimensionale Datenhaltung des Data Warehouse.

  Das ETL-Tool wird derzeit beschafft.

- **OLAP-Toolumgebung**

  Die OLAP-Toolumgebung dient zum Auswerten der Daten des Data Warehouse auf der Basis von vorgefertigten Berichtstrukturen oder durch „Ad-Hoc-Abfragen“. Die Ergebnisse werden wahlweise in tabellarischer und/oder grafischer Form aufbereitet. Darüber hinaus werden statistische Auswertungsmethoden angeboten.

  Die OLAP-Toolumgebung wird derzeit beschafft.

- **Metadaten-Tool**

  Das Metadaten-Tool ermöglicht dem Endanwender den Zugriff auf die Metadatenbasis, in der alle Komponenten und Ereignisse des Data Warehouse fachlich beschreiben sind. Das Metadaten-Tool ist Bestandteil eines Portals für den Zugang zum Data Warehouse.

  Das Metadaten-Tool wird derzeit beschafft.

4 Bisherige Ergebnisse

Auf der Grundlage der fachlich übergreifenden Datenbasis des Data Warehouse werden die Beschäftigtenstatistik im Data Mart BST und wichtige Arbeitsmarkt- und Beschäftigtendaten für das neue Informationssystem des IAB im Data Mart IAB bereitgestellt.
Den Grundstock der Beschäftigtendaten im Data Warehouse bilden die Meldungen zur Sozialversicherung nach der DEÜV seit 1999 sowie entsprechend aufbereitete DEVO/DÜVO-Meldungen ab 1997. Es werden monatsdurchschnittlich ca. 6 Mio. Beschäftigtendatensätze von den Rentenversicherungsträgern in einem neuen Satzformat (DEÜV) an die BA übergeben. Diese werden in terminlich festgelegten Ladeprozessen in der historisch korrekten Reihenfolge in der Datenbasis des Data Warehouse (DWH-Datenbasis) gespeichert. Derzeit ist der Datenbestand auf ca. 350 Mio. Beschäftigtendatensätze angewachsen. Zur Verarbeitung der Beschäftigtendaten gehören aber nicht nur die DEÜV-Meldungen, sondern auch ergänzende Daten, zu denen diese Meldungen in Beziehung gesetzt werden müssen (Dimensionsdaten). Dies sind Gebietsstrukturdaten (z.B. Orte, Gemeinden), Aufbauorganisationsdaten der BA (HSt – LAA – AA – GSt), Betriebedaten, Wirtschaftsklassifikationen und einige andere mehr. Alle Daten werden in der Regel monatlich oder bei Bedarf nach fachlich genau spezifizierten Vorgaben auf syntaktische und semantische Richtigkeit geprüft und je nach Prüfergebnis in die Datenbasis des Data Warehouse (DWH-Datenbasis) übernommen oder ausgesondert.

Auf der Grundlage dieser DWH-Datenbasis werden die quartalsweise fälligen Statistiken zu den sozialversicherungspflichtig und geringfügig Beschäftigten erstellt. Zum derzeitigen Berichtsprogramm gehören die Eingangs-, die Bestands-, die Betriebs- und Pendlerstatistiken, die Versorgung des IAB und des Statistischen Bundesamts mit Quartalseinzeldaten sowie die Belieferung anderer externer Stellen, insbesondere Forschungsinstitute. Die Beschäftigtenstatistik wurde vom zuständigen Fachreferat IIIa6 rückwirkend ab der Quartalsauswertung 30.6.1999 freigegeben. Die wichtigsten statistischen Tabellen werden auch in der Statistikdatenbank STADA aktualisiert.

Datentechnisches Fundament des neuen statistischen Analyse- und Informationssystems des IAB ist ebenfalls die DWH-Datenbasis des Data Warehouse der BA. Daten der Arbeitsmarktforschung werden aus unterschiedlichen, autorisierten Quellen übernommen und in der einheitlich strukturierten Datenbasis des Data Mart IAB zusammengeführt. Aufbereitet und qualitätsgeprüft stehen sie für Abfragen und Analysen unterschiedlicher Fachanwenderkreise, insbesondere der Regionalforschung, bereit.

5 Ausblick

Um künftig den Endanwendern die Möglichkeit zu bieten, die statistischen Ergebnisse direkt aus dem Data Warehouse abrufen zu können, wird die dafür erforderliche Software derzeit im Rahmen einer europaweiten Ausschreibung beschafft. Der Zuschlag wird im Dezember 2001 erteilt. Anschließend erfolgt die Integration in die Infrastruktur

der BA. Die Anpassung der Analysesoftware im Hinblick auf eine Nutzung im Bereich der Referate ICF der LAÄ und Sachgebiete IC der AÄ unter Beteiligung von Praktikern dieser Dienststellen und der Piloteinsatz in einigen ausgesuchten Dienststellen ist für das Jahr 2002 geplant.

Das Berichtsprogramm der Beschäftigtenstatistik wird ausgebaut. Die monatlichen Statistiken mit kürzeren Wartezeiten (bisherige 10%-Stichprobe) folgen – zukünftig auf Basis einer Totalauswertung aller Beschäftigtendaten – voraussichtlich im vierten Quartal 2001.

Die DWH-Datenbasis wird 2002 um weitere Daten erweitert werden. Als nächstes sollen statistische coArb-Daten und danach statistische coSach-Daten übernommen werden. Dabei wird das wesentliches Ziel der Datenkonsolidierung (Zuordnung der Daten zu den übergeordneten Personen- und Betriebedaten) verfolgt. Die Arbeitsmarktstatistik hat dann eine wesentlich verbesserte Datengrundlage. Das Berichtsprogramm der Arbeitsmarktstatistiken wird entsprechend den im Data Warehouse verfügbaren Daten neu organisiert.

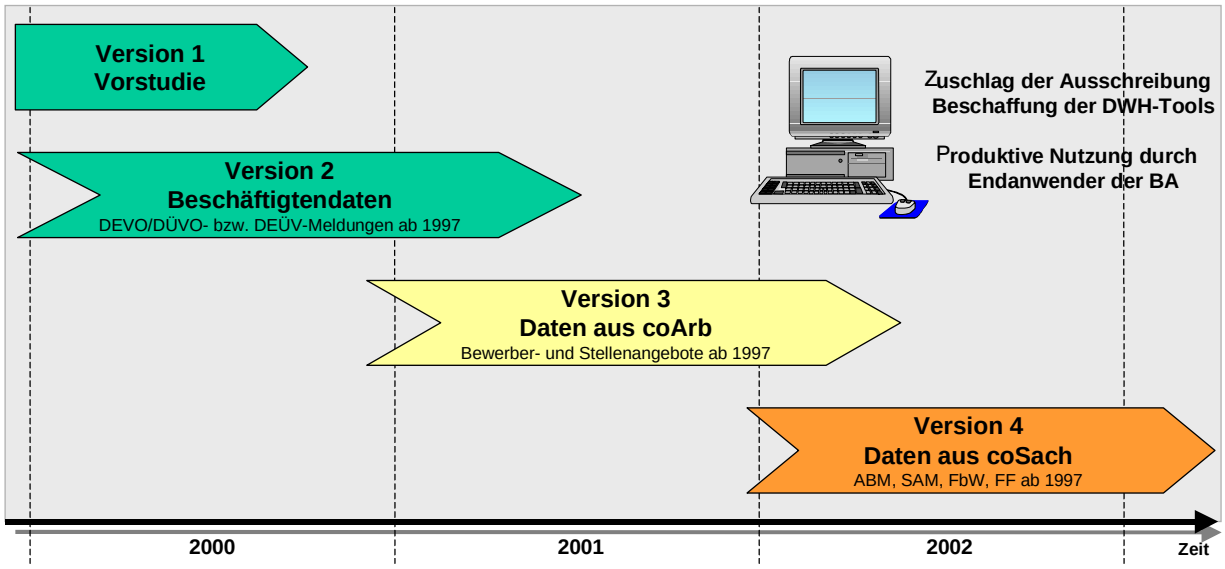


Bild 5: Planung des Data Warehouse der BA