Challenges in Large Enterprise Data Management

James Hamilton
JamesRH@microsoft.com
Microsoft SQL Server
2002.08.20
Enterprise Data Management Issues

- Three leading enterprise data management challenges:
  1. Availability/Cost of administration
  2. Legacy integration & multi-tier applications
  3. End-to-end system security

- Industry making progress but the job isn’t yet done

- Feature race partly diverts industry attention
1. Availability/Cost of Admin

- **1985 Tandem study (Gray):**
  - Administration: 42%, Software: 25%, Hardware 18% downtime
- **1990 Tandem Study (Gray):**
  - Software 62%, Administration: 15%
- **Even server-side S/W is big:**
  - Network Attached Storage: 1 mloc
  - Windows2000: Over 50 mloc
  - Database: 3 to 5 mloc
  - SAP: 37 mloc (4,200 engineers)
- **Example single outage costs** *(Patterson HPTS02 & InternetWeek 4/3/00)*:
  - Brokerage: $6,500k
  - Credit Card Auth: $2,600k
  - Ebay: $225k
  - Amazon.com: $180k
- **Observations:**
  - H/W downtime contribution trending to zero
  - Software & admin costs dominate & growing
  - Expensive: admin >5x S/W & H/W cost
  - Administration error prone: #1 or #2 cause of downtime
1. Availability/Cost of Admin

Solution summary:

- Understand downtime/measure improvements:
  - Event Log Analyzer
  - Watson
  - Data Collection Agent
- Ease of admin: “No knobs” design
- Auto-administration: Index tuning wizard
- Online utilities
- No-reboot diagnostics
  - Uncertain admins just restart
- Admin thread
  - Allow recovery without restart
2. Legacy Integration & Multi-tier Apps

Applications multi-tiered for many reasons:
- Wrap legacy systems as web services
- Improved availability through redundant mid-tier servers
- Application scaling & DB offload through caching

Multi-tiered applications typically still hand crafted
- Needed: object access layer, data cache, async queuing, data directed routing, mid-tier security support & integration, ...

Solution Summary:
- Visual Studio: App, mid-tier, & DB dev, deployment, & debug
- Distributed heterogeneous query support
- Direct HTTP access into database
- Multi-tier cache integration with notifications
- Semi-structured & unstructured search:
  - XML <> Relational Mapping
  - Native XML datatype support
  - XQuery & XPath support
  - High-scale full text indexing and unstructured search
- Asynchronous queuing in database
- Security integration without fully provisioning back-end DB
3. System Security

- **DB Security is as important as data integrity**
  - Protection against data damage, loss, manipulation, & unintentional disclosure

- **Rapidly changing threat environment:**
  - DBs on public net rather than behind mediated access (CS rep)
  - Customers & partners integrating & directly sharing data
  - Security “experts” built publishing potential vulnerabilities

- **Security issues are under reported**

- **Also under invested:**
  - “Less than 0.0025% of corp revenue invested in security” – Richard Clarke, Special security advisor to president

- **Solution summary:**
  - Full review of server software stack (~$100m investment)
    - Full SQL organization for nearly 3 months
  - Automation: Code analysis tools targeting security vulnerabilities
  - Address Admin error: Easy to understand, policy driven
  - Fix delivery: Installer & auto-update (300m downloads/month)
  - Research on active security:
    - Detecting new threats & miss-configurations