



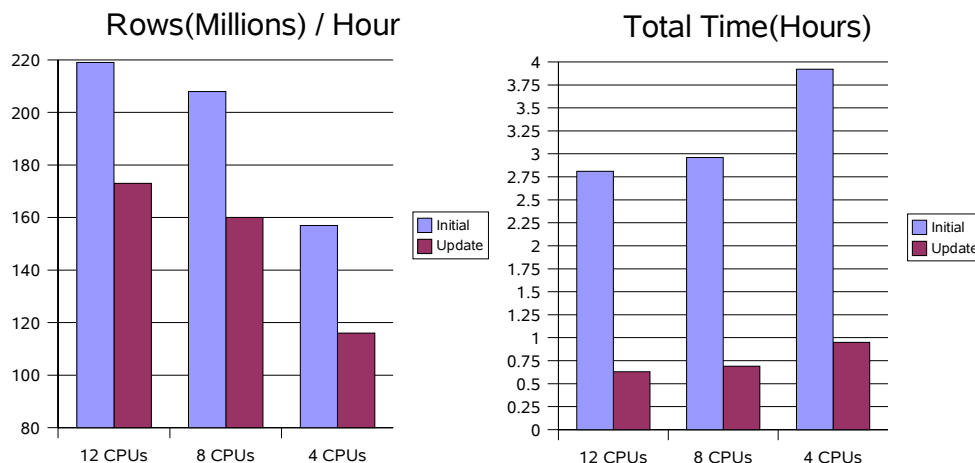
SAS® Enterprise ETL Server Performance Brief

Scalable Architecture, Solaris 10, Performance Leadership Out of the Box

Executive Summary

Configuration:

- Sun Fire E4900, 12/8/4 dual core CPUs¹
- Sun StorEdge 3510s
- Solaris 10
- SAS Enterprise ETL Server, 9.1.3, SP2
- SAS ETL Server Benchmark, v1.4
- Data Sizes:
 - Warehouse Load: 145GB, 616M rows
 - Warehouse Update: 29GB, 110M rows



Background

Utilizing the building block architecture of the Sun Scalable BIDW (Business Intelligence/Data Warehousing) platform in concert with the Solaris 10 Operating Environment, this performance brief summarizes a performance modeling exercise using the SAS ETL Server Benchmark, Version 1.4. The goal was to run the extra large data model in a scalable architecture fashion, namely, on a 12, 8, 4 dual core CPU configuration with a fixed storage configuration.

At any given time through this particular complex ETL process flow, there could be anywhere from 5 to 150+ processes running various segments of the workload, each of which are individually utilizing several multi-gigabyte files. Although the benchmark is dominated by an I/O intensive workload, complex interactions from

¹ UltraSPARC IV CPUs are dual core; psrinfo(1) will show 2 processors per CPU

the kernel, process scheduling and file system cache heuristics all come into play. Thus, this is the purpose of investigating the effects of varying the CPU configuration as it characterizes the true complexity and risk in any large scale application – not so much as how an individual application performs, but more importantly, to demonstrate its performance with all the other simultaneous tasks. As the results indicate, performance leadership is exemplified with all tested configurations.

Any given ETL flow could resemble this model closely or it may be completely different so care should be given to performance extrapolations from the given results.

SAS ETL Benchmark Scenario

At a high level, the test scenario consists of 2 phases which should be conceptually similar to many ETL flows.

Initial warehouse creation (Phase 1)

- Initial validation and load of raw data into warehouse store
- Transformation into enterprise defined schemas

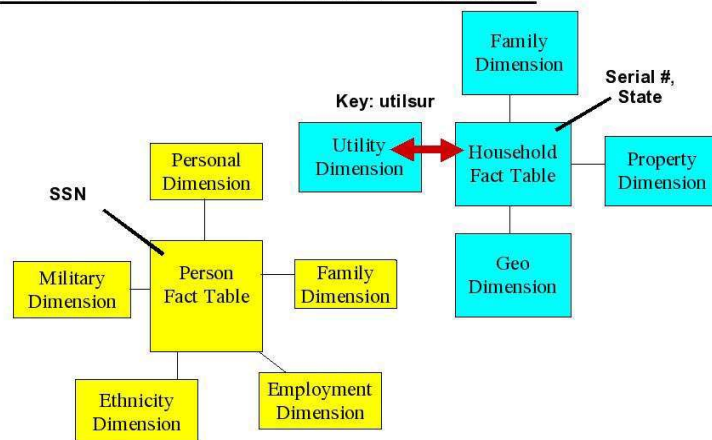
Periodic Warehouse Update (Phase 2)

- Same steps as above but typically with less data (~18% in our case); initial warehouse creation is not needed since it was already done in Phase 1

The SAS® ETL Server Benchmark test scenario builds a demographic data warehouse that includes details of both households and individuals using an augmented set of U.S. Census data. The first phase of the test builds out a data warehouse composed of two star schemas from the individual state Census data which is validated, loaded and then transformed into the star schema consisting of two fact tables (Person & Household) with associated dimensions. All dimension tables are related back to the fact table by a generated surrogate key which is built from multiple business values found in each dimension table. Data validation was performed on over 70% of the input values and includes tasks such as range checking, data conversion, simple calculations and string manipulation.

In a common data warehouse model it is typical to update or add 10-20% of the total warehouse records at regular, predefined intervals. The second phase of the test demonstrates *Slowly Changing Dimensions* (tracking of dimension table changes) during the update portion of the ETL process.

Data Model – Star Schema

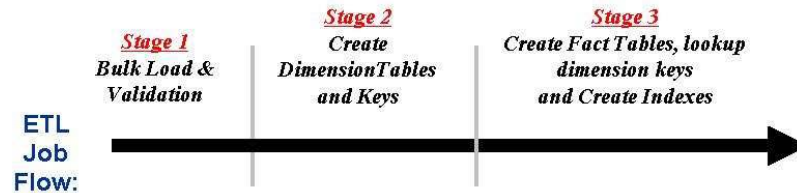


Indexes are created on dimension tables during the dimension create/update stage and fact tables created upon completion of the initial load and/or update process.

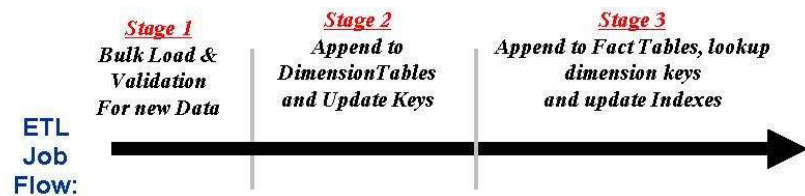
SAS ETL Server Benchmark Execution Highlights

Test Suite Execution Stages

Initial DW Build (First Time DW Creation):



Periodic DW Updates (example: Nightly ETL Load):



- Stage 1: Validate and load incoming data into warehouse
- Stage 2: Formation into star schema requiring creation of:
 - Dimension tables
 - Dimension table indexes
- Stage 3
 - Create fact tables
 - Surrogate key lookups via dimension tables
 - Index Creation

Test Results

The SAS ETL Server Benchmark is a complex transformation that imports data comprised of more than 100 variables or columns per row and loads into one of the two star schemas (household, personal).

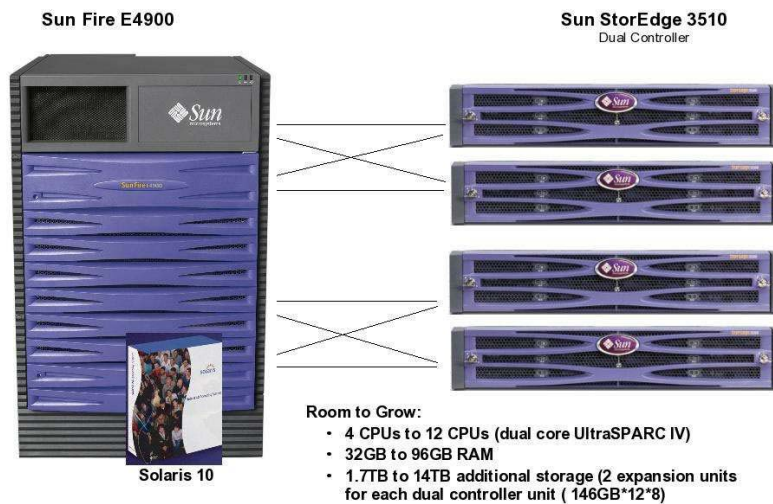
These impressive test results clearly demonstrate the ability to execute complex transformations on over 600 million rows in a small window of time. With the total number of rows processed, imagine loading 3 credit card transactions for every person in the United States in under 3 hours. This demonstrates that several complex ETL batch processes can execute simultaneously on any of the tested configurations with excellent performance results. Again, since the storage configuration remained fixed in this I/O intensive test, we do not expect scalability results linear to the number of CPUs.

	<i>12 CPUs</i>		<i>8 CPUs</i>		<i>4 CPUs</i>	
	Initial	Update	Initial	Update	Initial	Update
Totals						
Rows	616M	110M	616M	110M	616M	110M
Time	2.8hrs	.65 hr	2.9hrs	.70 hr	3.9hrs	.95 hr
Rows / Hr	219M	173M	209M	160M	157M	116M
GB / Hr	51.6	45.8	49.1	42.3	36.9	30.7

Table 1: SAS ETL Benchmark on Sun Fire E4900 results

Configuration Specifics

SAS Enterprise ETL Server Scalable Building Block Architecture



	Configuration	Notes
System	Sun Fire E4900	
	12 x 1.2 GHz US IV	12,8,4 dual core CPU configuration
	48GB RAM	
Storage	4 x Sun StorEdge 3510 FC 4 UFS data file systems (created with -m 1 -i 16384)	Dual controller units, 8 paths to the host dual pathed with Solaris Traffic Manager.
Software	Solaris 10, SAS 9.1.3 SP2, SAS ETL Server Benchmark Suite, v1.4	

The Sun Fire E4900 server is a large-scale shared memory system. Scaling up to 12 UltraSPARC IV processors, this server is ideal for server consolidation and optimized for running applications such as large departmental databases, customer management, decision support, as well as the most demanding ETL processes. For

maximum performance and availability, the Sun Fire E4900 server offers full hardware redundancy, fault-isolated Dynamic System Domains and Dynamic Reconfiguration.

The Sun StorEdge 3510 FC array uses a modular, building-block approach to help reduce costs and simplify future upgrades to entry-level SANs. Up to four dual path or eight single path servers can be connected to a dual controller tray without using a switch. This array is extremely easy to deploy, configure, manage, and monitor. With affordable enterprise-class features and functionality such as dual hot swap power and cooling, hot swap redundant RAID controllers, hot swap disk drives, global and local hot sparing, dynamic LUN expansion, dynamic capacity expansion, and remote status monitoring, the Sun StorEdge 3510 FC array is a standout products in its class and a natural component for the Sun Scalable BIDW platform.

Solaris 10 contains many features and performance optimizations and anchors the Sun Scalable BIDW platform. The Solaris 10 Operating Environment brings an unprecedented level of performance enhancements and innovation to the user with features such as Solaris Containers, Dtrace, ZFS, Process Rights Management, and Predictive Self Healing. But the decision to upgrade to Solaris 10 from earlier Solaris releases is worry free with Sun Microsystems' guarantee for application binary compatibility. This guarantee preserves customer investments like no other Operating System can match.

Summary

Performance Excellence at each configuration

These results speak clearly to the performance advantages of running the SAS Enterprise ETL Server on Solaris 10. An orchestration of application, OS and select HW configuration creates a symphony for IT departments responsible for deployment. Performance leading results were achieved with each configuration despite the benchmark being heavily dominated by an I/O centric workload. The Sun Fire E4900 configuration in conjunction with Solaris 10 can handle the most rigorous and demanding applications. In today's enterprise, the only constant is change. Application deployment on this platform can be considered a safety net or an insurance policy as the system performs predictably and reliably under heavy workloads. Knowing that performance is so robust and solid for such a demanding set of tasks tremendously mitigates the risk from the unknown demands of next month or quarter. The real time re-tooling of business processes require rapid response to shifts in priorities and resource allocations. Choosing the right BI/DW platform is critical to the dynamic nature of today's business.

Out of the Box – As Is!

And if these performance leading results aren't impressive enough, **no** system optimizations, file system or kernel tuning was done! This was Solaris 10 out of the box (installable in just a few short minutes). Not many systems administrators have the time or necessarily the expertise to modify system and/or kernel variables, run extended performance testing and lengthy what-if scenarios . Simplicity, manageability, predictability, observability and stability are paramount. Solaris 10 – it just works out of the box.