Accelerating Data Science Workflows: Gen-over-Gen CPU Gains on Intel Processors

Planning / Implementation

Most enterprise analytics pipelines still lean on Python DataFrames (pandas/Modin) and classical ML libraries (scikit-learn, XGBoost), so CPU-only efficiency directly impacts cost, latency, and throughput. As Intel Xeon generations advance, pairing Modin and Intel Extension for Scikit-learn turns architectural gains into real end-to-end time savings with minimal code change.

todo: what is this list? add an intro statement

- Pandas ↔ Modin. Pandas is the de-facto DataFrame API; Modin keeps that API while parallelizing execution across cores/cluster back-ends (Ray). This allows parallel I/O and compute with minimal code change (import swap).
- Intel Extension for Scikit-learn (sklearn-intelex). A single call to patch_sklearn() dynamically patches popular estimators to highly-optimized C++ kernels (oneDAL), accelerating both fit and predict without rewriting pipelines.
- Xeon generations. We focus on realistic CPU-only deployments comparing 3rd Gen Xeon Scalable, 5th Gen Xeon, and 6th Gen Xeon ("Xeon 6").

Series of papers

This paper represents Part 3 of a series on Accelerating Data Science Workflows:

- Part 1: Modin vs. Pandas for data manipulation (I/O, transforms)
- Part 2: Intel Extension for Scikit-learn (training & inference)
- Part 3 (this paper): End-to-end gains across 3rd, 5th, and 6th Gen Intel Xeon CPUs, covering data manipulation + model training + inference, with composite pipeline speed-ups

Algorithms and Datasets

To ensure apples-to-apples comparisons, we reuse the workloads from Parts 1–2 and run them on identical software stacks across 3rd, 5th, and 6th Gen Intel Xeon CPUs. The subsections below specify workloads, metrics, and environment details. todo: which subsections? do you just mean the text below?

- Workloads:
  - Data manipulation (Modin): CSV ingest + transforms at 400K / 800K / 1.6M rows.
  - Training (sklearn-intelex): DBSCAN, K-means, KNeighborsClassifier, Logistic/Linear Regression, Random Forest Classifier/Regressor.
  - Inference: CatBoost, LightGBM, XGBoost, plus the scikit-learn models above.
- Metrics: Wall-clock time for each operation; composite results use median to reduce skew from outliers. We also report 6th↔5th and 6th↔3rd speed-up factors.
- Environment: Same software stack across generations; CPU generations as titled. todo: what does this mean "titled"?

Methodology Notes & Reproducibility

To keep results fair and repeatable, we standardize seeds, force garbage collection between iterations, and repeat runs to smooth variance. Use the notes below to replicate our setup and verify the numbers.

- All timings are wall-clock.
- Each test repeated multiple iterations with garbage collection (gc.collect()) between runs; median reported.
- Data manipulation used Modin; ML used scikit-learn patched with Intel Extension for Scikit-learn.
- Keep the same software version, BIOS settings, and dataset descriptions.

Results

In this section, we list the raw wall clock times and normalized speed ups (5th Gen to 6th Gen, and 3rd Gen to 6th Gen). Lower values are better for time; The ratios in the right two columns highlight 6th Gen uplifts.

- Data Manipulation (Modin)
- Model Training (Intel Extension for Scikit learn)
- Model Inference (Intel Extension for Scikit learn)
- Composite “Full Pipeline” View

todo: would "Improvement" be a better word that "Uplift" in these tables?

Data Manipulation (Modin)

The improvement for Modin across the three sizes are shown in the following table:

- 6th vs 5th = 1.88×-2.64×
- 6th vs 3rd = 3.00×-6.13×

todo: what is the (s) in the heading row - eg "3rd Gen (s)"?

Table 1. Data Manipulation (Modin)

Rows	3rd Gen (s)	5th Gen (s)	6th Gen (s)	6th/5th Uplift ↑	6th/3rd Uplift ↑
400,000	21.80	13.64	7.26	1.88×	3.00×
800,000	39.97	17.36	8.66	2.00×	4.62×
1,600,000	60.05	25.85	9.80	2.64×	6.13×

Model Training (Intel Extension for Scikit learn)

Across algorithms (training): todo: reword this intro to refer to the table

- 6th vs 5th ranges 1.11×–3.00×
- 6th vs 3rd ranges 1.44×–7.23×

Table 2. Model Training (Intel Extension for Scikit learn)

Algorithm	3rd Gen	5th Gen	6th Gen	6th/5th Uplift ↑	6th/3rd Uplift ↑
DBSCAN	9.921	6.862	5.080	1.35×	1.95×
K means	1.295	1.069	0.558	1.92×	2.32×
KNeighborsClassifier	1.257	1.016	0.609	1.67×	2.06×
Linear Regression	0.020	0.015	0.005	3.00×	4.00×
Logistic Regression	22.632	17.541	15.748	1.11×	1.44×
Random Forest Classifier	2.678	0.682	0.402	1.70×	6.66×
Random Forest Regressor	59.208	12.161	8.188	1.49×	7.23×

Model Inference (Intel Extension for Scikit learn)

Across algorithms (inference): todo: reword this intro to refer to the table below

- 6th vs 5th ranges 1.19×–2.38×
- 6th vs 3rd ranges 1.63×–4.37×

Table 3. Model Inference (Intel Extension for Scikit learn)

Algorithm	3rd Gen	5th Gen	6th Gen	6th/5th Uplift ↑	6th/3rd Uplift ↑
CatBoost	0.0031	0.0027	0.0017	1.59×	1.82×
K means	0.0050	0.0038	0.0016	2.38×	3.13×
KNeighborsClassifier	1.3968	0.4396	0.3194	1.38×	4.37×
LightGBM	0.0030	0.0016	0.0007	2.29×	4.29×
Linear Regression	0.0019	0.0014	0.0008	1.75×	2.38×
Logistic Regression	0.1431	0.1169	0.0708	1.65×	2.02×
Random Forest Classifier	0.4425	0.2854	0.2044	1.40×	2.16×
Random Forest Regressor	0.4409	0.1640	0.1097	1.49×	4.02×
XGBoost	0.0026	0.0019	0.0016	1.19×	1.63×

Composite “Full Pipeline” View

To avoid over-weighting any single stage, we compute a balanced median uplift across the three segments (Modin data manipulation, training, inference):

- 6th vs 5th Gen: ≈ 1.6x – 2.5x
- 6th vs 3rd Gen: ≈ 2.5x – 6.1x

Note*:* If your workload is training-heavy or inference-heavy, scale each segment by the appropriate share of wall-clock time to obtain the scenario-specific uplift.

Conclusions

6th Gen Intel Xeon consistently advances end-to-end, CPU-only analytics versus 5th and 3rd Gen baselines using the same, familiar software stack (Modin + Intel Extension for Scikit-learn). In our tests, 6th/5th uplifts typically land in the 1.88×–2.64× range for data manipulation, 1.11×–3.00× for model training, and 1.19×–2.38× for inference; against 3rd Gen, ranges widen to 3.00×–6.13×, 1.44×–7.23×, and 1.63×–4.37×, respectively. The full-pipeline gain is workload-dependent but commonly falls around ≈1.6×–2.6× vs 5th Gen (and ≈2.5×–6.1× vs 3rd Gen) when combining prep, fit, and predict.

Practical takeaways:

- **Prioritize CPU upgrades** when your pipelines are **I/O-heavy** (large CSV/Parquet ingestion, wide group-bys) or rely on **tree ensembles, clustering, or distance-based methods** . These show the largest improvements from gen-to-gen and from the Intel-optimized kernels.
- **Mind training vs inference trade-offs.** Some algorithms may train modestly faster but infer dramatically faster (or vice-versa). Choose CPU generation and algorithmic settings based on where your SLA or cost is constrained (e.g., batch-training windows vs. online latency).
- **Adoption is low-friction.** The improvements arrive with **minimal code change**: an **import swap** for Modin and a **patch_sklearn()** call for Intel Extension for Scikit-learn, preserving APIs and model semantics.
- **Right-size with pipeline weights.** Apply the per-stage ranges to your own time profile (e.g., 40% prep / 35% train / 25% infer) to estimate business impact. Where inference dominates, favor gains in predict-time algorithms; where training windows dominate, weight fit-time uplifts more heavily.

Overall, upgrading to 6th Gen Intel Xeon turns many formerly multi-second steps into sub-second operations and materially compresses end-to-end latency, without abandoning the mainstream pandas/scikit-learn ecosystem.

Lab Configurations

Our test server had the hardware and software configuration listed in the following table.

todo: I note that the memory is different for each server. Does this make a difference? How about adding some text about this?

Table 4. Lab Configurations

Component	3rd Gen Intel Xeon Server	5th Gen Intel Xeon Server	6th Gen Intel Xeon Server
Server configuration			
Platform	Lenovo ThinkSystem SR650 V2	Lenovo ThinkSystem SR650 V3	Lenovo ThinkSystem SR650 V4
CPU	Intel Xeon 4310 processor, 24 cores / 48 threads @ 3.0 GHz	Intel Xeon 8592+ processor, 64 cores / 128 threads @ 3.9 GHz	Intel Xeon 6787P processor, 86 cores / 172 threads @ 3.8 GHz
Memory	16x 16 GB DDR4 RAM	16x 32 GB DDR5 RAM	16x 64 GB DDR5 RAM
OS	Ubuntu 22.04.5 LTS (Linux kernel 6.8.0-59-generic)	Ubuntu 22.04.5 LTS (Linux kernel 6.8.0-59-generic)	Ubuntu 22.04.5 LTS (Linux kernel 6.8.0-59-generic)
Software components			
Python	Version 3.10.12		
Pandas	Version 2.2.3		
scikit-learn	Version 1.5.0		
scikit-learn-intelex	Version 2025.1		

References

For more information, see these web resources:

- Modin documentation
  https://modin.readthedocs.io/
- Intel Extension for Scikit-learn Overview:
  https://www.intel.com/content/www/us/en/developer/tools/oneapi/scikit-learn.html ;
- Intel Extension for Scikit-learn Getting Started:
  https://www.intel.com/content/www/us/en/developer/articles/guide/intel-extension-for-scikit-learn-getting-started.html
- 3rd Gen Intel Xeon Scalable:
  https://www.intel.com/content/www/us/en/products/docs/processors/xeon-accelerated/3rd-gen-xeon-scalable-processors.html
- 5th Gen Intel Xeon Scalable:
  https://www.intel.com/content/www/us/en/products/docs/processors/xeon/5th-gen-xeon-scalable-processors.html
- Intel Xeon 6:
  https://www.intel.com/content/www/us/en/products/details/processors/xeon.html

Authors

Kelvin He is an AI Data Scientist at Lenovo. He is a seasoned AI and data science professional specializing in building machine learning frameworks and AI-driven solutions. Kelvin is experienced in leading end-to-end model development, with a focus on turning business challenges into data-driven strategies. He is passionate about AI benchmarks, optimization techniques, and LLM applications, enabling businesses to make informed technology decisions.

David Ellison is the Chief Data Scientist for Lenovo ISG. Through Lenovo's US and European AI Discover Centers, he leads a team that uses cutting-edge AI techniques to deliver solutions for external customers while internally supporting the overall AI strategy for the Worldwide Infrastructure Solutions Group. Before joining Lenovo, he ran an international scientific analysis and equipment company and worked as a Data Scientist for the US Postal Service. Previous to that, he received a PhD in Biomedical Engineering from Johns Hopkins University. He has numerous publications in top tier journals including two in the Proceedings of the National Academy of the Sciences.

Notices

Lenovo may not offer the products, services, or features discussed in this document in all countries. Consult your local Lenovo representative for information on the products and services currently available in your area. Any reference to a Lenovo product, program, or service is not intended to state or imply that only that Lenovo product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any Lenovo intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any other product, program, or service. Lenovo may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

Lenovo (United States), Inc.
8001 Development Drive
Morrisville, NC 27560
U.S.A.
Attention: Lenovo Director of Licensing

LENOVO PROVIDES THIS PUBLICATION ”AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. Lenovo may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

The products described in this document are not intended for use in implantation or other life support applications where malfunction may result in injury or death to persons. The information contained in this document does not affect or change Lenovo product specifications or warranties. Nothing in this document shall operate as an express or implied license or indemnity under the intellectual property rights of Lenovo or third parties. All information contained in this document was obtained in specific environments and is presented as an illustration. The result obtained in other operating environments may vary. Lenovo may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any references in this publication to non-Lenovo Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this Lenovo product, and use of those Web sites is at your own risk. Any performance data contained herein was determined in a controlled environment. Therefore, the result obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

This document, LP2269, was created or updated on September 17, 2025.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP2269
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP2269.

Trademarks